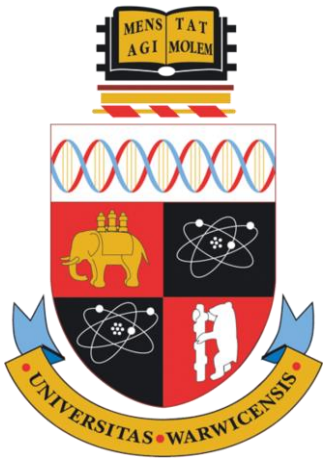


Optimising the FIT: Risk adjusted Colorectal Cancer Screening using Routine Data

Jennifer Anne Cooper MSc BSc

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the Health Sciences

The University of Warwick,
Warwick Medical School, Division of Health Sciences

April 2018

"The past, like the future, is indefinite and exists only as a spectrum of possibilities."

– Stephen Hawking

Contents

Preface

Chapter 1: Introduction

Chapter 2: Systematic Review of Risk Prediction Models Combining the FIT Result for Colorectal Cancer Screening

Chapter 3: Risk-adjusted Colorectal Cancer Screening Using the FIT and Routine Data: Development of a Risk Prediction Model

Chapter 4: Development of a Risk Prediction Model for Colorectal Cancer Screening using an Artificial Neural Network

Chapter 5: Investigating the Use of Routine GP Patient Data to Improve Colorectal Cancer Screening Referral Decisions

Chapter 6: THIN Data Extraction Methodology

Chapter 7: Thesis Discussion

Acknowledgements

Firstly, I would like to express my sincere gratitude towards my supervisors Dr. Sian Taylor-Phillips, Dr. Nick Parsons (for showing me the ways of R), Dr. Chris Stinton and Dr. Steve Smith for all their guidance over the last few years. I also thank Karoline Freeman for her inspiring input over the final few months.

I would like to thank Professor Tom Marshall for allowing me the opportunity to carry out an NIHR Doctoral Exchange at Birmingham University and Dr. Ronan Ryan for his time spent helping me to unravel the THIN database.

I would also like to acknowledge Professor Stephen Halloran for his support, in-depth knowledge and unwavering enthusiasm during my PhD and Andrew Taylor for helping me to recognise intangible progress.

Finally, Max T. Taberham for his support and being my light at the end of the tunnel.

Declaration

All material contained in this thesis is the candidate's own work. The thesis has not been submitted for a degree at another university.

The author has published the following articles related to the research:

Cooper, J. A., et al. (2018). Development of risk prediction models combining routine EHR data for use in colorectal cancer screening referral decisions [abstract]. In: Methods for Evaluating Medical Tests and Biomarkers: Utrecht, the Netherlands. 2-3 July 2018. Diagnostic and Prognostic Research 2018, 2(Suppl 1):P12.

Cooper, J. A., et al. (2018). "Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model." Br J Cancer 118(2): 285-293.

Cooper, J. A., et al. (2017). Risk-adjusted colorectal cancer screening using the FIT: development of a risk prediction model [abstract]. In: Methods for Evaluating Medical Tests and Biomarkers: Birmingham, UK. 19–20 July 2016. Diagnostic and Prognostic Research 2016, 1(Suppl 1):P1.

Cooper, J. A., et al. (2016). "FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test." Colorectal Dis **18**(7): 650-653.

Further publications and collaborative work:

Seedat F., et al. (2014). "International comparisons of screening policy-making: A systematic review." National Screening Committee.

Tsertsvadze, A., et al. (2016). "Community-onset sepsis and its public health burden: a systematic review." Syst Rev **5**: 81.

Taylor-Phillips, S., et al. (in progress). "Effect of Computer-Aided Detection Prompts on Breast Screening Performance."

Jennifer Anne Cooper is supported by the NIHR CLAHRC West Midlands initiative. This thesis presents independent research and the views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

This thesis explores the value of risk-adjusted colorectal cancer screening using the faecal immunochemical test (FIT). Following the FIT pilot study in the English Bowel Cancer Screening Programme (BCSP), there was opportunity to investigate a risk-adjusted approach to screening. This thesis was informed by several evolving areas of research including risk prediction modelling, test accuracy, as well as the use of electronic health records (EHRs) and the statistical methods best applied to utilise these data. The emphasis of the research was on routine data and used the Bowel Cancer Screening System (BCSS) as well as anonymised GP records for model development (THIN – The Health Improvement Network). Three statistical modelling techniques were investigated to build a risk prediction model for use in screening referral based decisions. A conventional approach using logistic regression was investigated first, which showed an improvement in both model performance and test accuracy for the risk-adjusted model over the FIT alone. This model was then extended further by investigating a machine learning algorithm in the form of an artificial neural network. An advantage of this approach is the flexibility to model complex nonlinear associations. The performance of this model (discrimination), as well as the sensitivity when applied as a test, was significantly better than the logistic regression model. Next an anonymised GP record was investigated for additional predictors to add to a risk-adjusted model using survival analysis, which exploits the longitudinal nature of the data. Two models were produced; one which combined the faecal occult blood test (FOBT) with lab test results, symptoms and other predictors and another which was developed for those with negative FOBT results only, to determine whether additional predictors could be used for referral decisions. In order to utilise EHRs for research, the methods need to be reproducible and transparent. An Acceptable Electronic BCSP (AEB) date was developed for quality assurance of the primary care data as well as to help determine a screening cohort for analysis. As a collective, these studies show evidence for improved performance of risk-adjusted screening over using the FIT alone. Future research should focus on further BCSS predictors and external validation of a risk-adjusted model in the BCSP. Machine learning approaches may be better placed for more complex electronic data. Future risk prediction model studies should encompass the whole pathway from model development to external validation and model impact before being implemented in practice with the ultimate aim of improving patient outcomes.

Abbreviations

Abbreviation	Description
AA	Advanced Adenoma
ACU	Acceptable Computer Usage
AEB date	Acceptable Electronic BCSP Date
AFT	Accelerated Failure Time
AHD	Additional Health Data
AIC	Akaike Information Criterion
AIS	Additional Information Services
AMR	Acceptable Mortality Reporting
ANN	Artificial Neural Network
ATC	Anatomical Therapeutic Chemical Classification System
AUC ROC	Area under the Receiver Operating Characteristic Curve
BCSP	Bowel Cancer Screening Programme
BCSS	Bowel Cancer Screening System
BIC	Bayes Information Criterion
BLR	Bayes Logistic Regression
BMI	Body Mass Index
BNF	British National Formulary
BSREC	Biomedical and Scientific Research Ethics Committee
CALB	Calgranulin B
CHARMS	Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (Checklist)
CI	Confidence Interval
CLAHRC	Collaborations for Leadership in Applied Health Research and Care
COGENT	Colorectal cancer GENeTics
CPRD	Clinical Practice Research Datalink
CRC	Colorectal Cancer
DF	Degrees of Freedom
EHR	Electronic Health Records
FAP	Familial Adenomatous Polyposis
FIT	Faecal Immunochemical Test
FN	False Negative
FOBT	Faecal Occult Blood Test
FP	False Positive
FS	Flexible Sigmoidoscopy
GP	General Practitioner
GRIPS	Genetic Risk Prediction Studies (Reporting Guidelines)
Hb	Haemoglobin
HNPCC	Hereditary Non Polyposis Colorectal Cancer
HR	Hazard Ratio
HTA	Health Technology Assessment
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
ICD	International Classification of Diseases
IMD	Index of Multiple Deprivation
KM	Kaplan Meier
LL	Log Likelihood
LOWESS	Locally Weighted Scatterplot Smoothing
LR	Logistic Regression
MAR	Missing at Random
MCV	Mean Cell Volume

MFP	Multivariable Fractional Polynomials
MLE	Maximum Likelihood Estimate
MNAR	Missing Not At Random
MSE	Mean Squared Error
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NPV	Negative Predictive Value
NRI	Net Reclassification Improvement
NSC	National Screening Committee
ODR	Office for Data Release
OR	Odds Ratio
PH	Proportional Hazards
PICOTS	Participants, Intervention, Comparator, Outcome, Timing, Setting
PMIP	Pathology Messaging Implementation Project Messaging
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction study Risk Of Bias ASsessment Tool
PROGRESS	PROGnosis RESearch Strategy
QOF	Quality and Outcomes Framework
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies
RCT	Randomised Controlled Trial
RECORD	REporting of studies Conducted using Observational Routinely-collected health Data (Statement)
ROC	Receiving Operating Characteristic Curves
SD	Standard Deviation
SRC	Scientific Review Committee
SSE	Sum of Squared Errors
SSP	Specialist Screening Practitioner
STARD	STAndards for the Reporting of Diagnostic accuracy studies (Reporting Guidelines)
STROBE	STrengthening the Reporting of OBservational studies in Epidemiology (Reporting Guidelines)
THIN	The Health Improvement Network
TN	True Negative
TP	True Positive
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (Reporting Guidelines)

List of Tables

Chapter 2: Systematic Review of Risk Prediction Models Combining the FIT for Colorectal Cancer Screening

Chapter 3: Risk-adjusted Colorectal Cancer Screening Using the FIT and Routine Screening Data: Development of a Risk Prediction Model

Chapter 4: Development of a Risk Prediction Model for Colorectal Cancer Screening using an Artificial Neural Network

Chapter 5: Investigating the Use of Routine Patient Data to Improve Colorectal Cancer Screening Referral Decisions

Chapter 6: THIN Data Extraction Methodology

List of Figures

Chapter 2: Systematic Review of Risk Prediction Models Combining the FIT for Colorectal Cancer Screening

Chapter 3: Risk-adjusted Colorectal Cancer Screening Using the FIT and Routine Screening Data: Development of a Risk Prediction Model

Chapter 4: Development of a Risk Prediction Model for Colorectal Cancer Screening using an Artificial Neural Network

Chapter 5: Investigating the Use of Routine Patient Data to Improve Colorectal Cancer Screening Referral Decisions

Chapter 6: THIN Data Extraction Methodology

Optimising the FIT: Risk adjusted Colorectal Cancer Screening using Routine Data

The introduction is based in part on the following editorial publication: Cooper, J. A., et al. (2016). "FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test." Colorectal Dis **18**(7): 650-653.

1.0 General Introduction

The lifetime risk of developing some form of cancer has increased to over 50% (1 in 2) for people born after 1960.[1] Worldwide, colorectal cancer (CRC) is the 3rd most common cancer in men and the 2nd most common in women with over half the cases occurring in more developed regions including North America, Australia, and Europe.[2] There were around 694,000 deaths from colorectal cancer globally in 2012.[2] In the UK, there were 41,804 new cases in 2015 and 15,903 deaths in 2014.[2] The lifetime risk in the UK for developing CRC in 2012 was 1 in 14 for men and 1 in 19 for women.[3] In 1975, the lifetime risk of CRC was 4% this has now increased to 6% in females and 7% in males in 2010.[4]

Public health screening is a process to identify individuals who may be at increased risk of a condition or disease.[5] Further investigation, treatment and information can be offered to these individuals to reduce their risk of developing the disease or to minimise complications that may arise.[5] A further definition by Raffle and Gray is summarised in the Box below and emphasises the focus of screening on the programme and not solely the screening test to identify these individuals.[6] A report published in 1968 for the World Health Organisation (WHO) by Wilson and Jungner defined ten principles of screening.[7] These principles have had a lasting influence on policy making for screening programmes with many countries using these principles or variations of them when deciding whether to implement a screening programme.[8] These principles have evolved into the 20 criteria which the UK National Screening Committee (NSC) use to appraise the viability, effectiveness and appropriateness of a screening programme.[9] These include aspects which relate to the condition, the test, the intervention, the screening programme and implementation criteria.

Raffle and Gray[6] Definition of Screening (page 37):

- "Testing of people who either do not have or have not recognised the signs or symptoms of the condition being tested for. In other words, they believe themselves to be well in relation to the disease that the screening relates to.
- Where the stated or implied purpose is to reduce risk for that individual of future ill health in relation to the condition being tested for, or to give information about risk that is deemed valuable for that individual even though risk cannot be altered.
- It encompasses the whole system or programme of events necessary to achieve risk reduction. Screening is a programme not a test" [6] (p. 37)

In 2007, Hewitson *et al.* showed that bowel cancer screening could reduce mortality from CRC by 16% in a meta-analysis of 4 randomised trials.[10] The National Screening Committee (NSC) recommended bowel cancer screening using a guaiac faecal occult blood test (gFOBT) following a pilot study for men and women aged 50 to 74 in July 2003.[11] Currently in England, men and women aged 60 to 74 are offered a biennial gFOBT.[12] A more recent test called the faecal immunochemical test (FIT) has been shown to have superior accuracy than the gFOBT.[13 14] This test has several other advantages over the gFOBT including the requirement of a single stool sample, greater specificity for human haemoglobin and an adjustable haemoglobin concentration threshold. The haemoglobin concentration detected by the test has also been shown to relate to risk of colorectal cancer,[15] and severity of colorectal cancer lesions.[16 17] As a result the FIT has been recommended for CRC screening by the *European guidelines for quality assurance in CRC screening and diagnosis* in 2010.[18] Many countries have now adopted FIT screening including Australia, New Zealand, the Netherlands, France, Canada, Spain, Italy and the Republic of Ireland.[19 20] The US Preventative Services Task Force recommend several screening strategies including the FIT for those at average risk aged 50-75 years.[21]

A 6 month pilot study of FIT versus the current guaiac FOBT was carried out in April 2014 by the NHS Bowel Cancer Screening Programme (BCSP) in England.[22] This study found a statistically significant greater uptake of screening with the FIT compared to gFOBT (66.4% versus 59.3%), and increased cancer and advanced adenoma detection rate. At an internationally used threshold of 20µg/g, the FIT positivity was 7.8% compared to the equivalent gFOBT positivity of 1.7%. Increasing the threshold to 180 µg/g gave a FIT positivity of 1.52%. Due to the improved uptake and higher test positivity when using internationally derived thresholds, extra demand would be put on a limited colonoscopy resource. The pilot study suggests there would be around 290,000 additional participants a year. This number would not be manageable for the current colonoscopy service. To ensure that capacity can be met, alternative thresholds for positivity (between 150-180 µg/g) are being investigated.

A suggested threshold to match colonoscopy resource and to give a similar number of people recalled to gFOBT was 160 µg/g. The NHS BCSP in England plans to adopt the FIT by the end of 2018.

In addition to amending the threshold of the FIT, stratified/personalised risk based CRC screening could be implemented to improve effective colonoscopy use, test accuracy and consequently health outcomes.[22-24] Stratified medicine aims to identify patients who would have a greater clinical benefit or least harm from a specific treatment.[25] A few studies have developed risk prediction models which combine the FIT concentration with other risk indicators for use in screening referral decisions.[15 24 26 27] This approach reserves the expensive and invasive colonoscopy resource for those at higher risk as estimated by a risk prediction model. By obtaining absolute risk predictions for individuals, those at higher risk can be referred on for diagnostic testing and those at lower risk can be placed back into the screening pool for continued surveillance every two years. This approach may optimise the benefits and harms of screening as well as available resource use.

The risk prediction model development study by Stegeman *et al*[26] combined the FIT result with risk indicators collected from a lifestyle questionnaire in a multivariable logistic regression model. The model had greater discrimination, with an area under the ROC curve of 0.76 than FIT alone (0.69). Test accuracy parameters were estimated with the risk based model which had a sensitivity of 40% compared to 32% for the FIT alone (at 93% specificity). The study by Tao *et al*[28] combined blood based inflammatory markers with the FIT and showed improved discriminatory power when comparing FIT alone (AUC 0.683) to a model combining FIT with 3 blood based markers (AUC 0.729). Test accuracy also showed improved sensitivity for the multivariable model.

Studies which require additional lab testing or the completion of a questionnaire have been shown to decrease screening uptake.[29] Electronic Health Records (EHRs) used in screening programmes and for healthcare in the UK are a rich resource of routine data and are being increasingly utilised in research. The NHS Bowel Cancer Screening Programme (BCSP) use the Bowel Cancer Screening System (BCSS) to record information for participants invited to screening. Information stored on this system includes invitation dates, whether an individual returned a screening kit, the result of the screening test, attendance at a SSP (Specialist Screening Practitioner) clinic and colonoscopy or diagnostic test results. Other EHRs include

those used by primary care to record diagnoses and symptoms as well as lifestyle factors, routine lab tests, drug prescriptions and anthropometrics. The interconnectivity of these different health systems in the NHS could be exploited for future research. By using routine data available from EHRs, the data are subject to quality assurance standards that improve data accuracy and completeness, as well as reducing participant burden. Furthermore, prediction models which utilise predictors available in routine care allow for greater application of the model in clinical practice.[30]

The area of risk prediction models or prognosis research compared to intervention studies and even diagnostic accuracy studies is an evolving area of research. There is growing interest in prediction model studies as reflected in the number of publications in recent years.[30 31] A prognostic model is described as 'a formal combination of multiple predictors from which risks of a specific end point can be calculated for individual patients'.[30] Prediction model research encompasses both prognostic models, which estimate the absolute probability that a certain outcome will occur within a specific time period in an individual, and diagnostic prediction models, which estimate the absolute probability that a certain outcome is already present.[32] The research in this thesis considers mainly the latter type. More guidance to improve research quality in this area has been published in recent years with the establishment of a Cochrane Prognosis Methods Group (PMG) (http://methods.cochrane.org/prognosis/welcome) and the PROGRESS (PROGnosis RESearch Strategy) research group (www.progress-partnership.org). The PROGRESS strategy series sets out a framework of 4 interconnected research themes for prognosis research and provides evidence to improve current research standards. Several guidelines, checklists and quality appraisal tools have been recently published for prognostic and risk prediction studies.[33-36] These are described in greater detail in the Systematic Review (**Chapter 2**).

Risk prediction models can be used in a variety of ways in healthcare including for guiding referral and treatment decisions as well as follow up strategies, cost effectiveness analyses and for shared decision making.[30] Although prediction models can provide a risk score, often this is used to assign individuals to risk groups and therefore the estimate of absolute probability is more accurate. This thesis focuses on producing risk prediction models which provide an absolute risk probability for use in decision making. The application of prediction models also relates to the shift towards personalised or stratified medicine where treatment plans and referral decisions are based on an individual's covariate (predictor) pattern. Since

prognostic and prediction model research, if applied on a population can lead to a change in health outcomes, and possibly to the spectrum of diagnosed disease, they can be considered a type of health technology assessment.[37 38] This is where the crossovers between applying a prediction model as a test can occur and is a key consideration of the thesis.

Risk based screening can be implemented at different points within the screening pathway. For instance, risk stratification of the target population to decide who should undergo the FIT could be performed to identify those at highest risk. This approach is investigated by Aniwan *et al*. using the Asia-Pacific Colorectal Screening System (APCS) as a preselection/triage before applying the screening test.[39] Alternatively, risk predictors could be integrated within the screening algorithm at the time of screening to decide who to refer on for colonoscopy.[26] Finally post screening surveillance strategies could be tailored by individual FIT results.[40 41] Timing is an important consideration of the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist, 'intended moment of using the model'.[33] This thesis focuses upon the combination of risk information around the time of the screening test (before colonoscopy) to determine high risk individuals to refer on for further diagnostic testing.

Different statistical methods can be used to build a prediction model, the most commonly used methods are regression, but other machine learning approaches can also be used such as neural networks, support vector machines, classification trees and random forests.[36] In this thesis, three different statistical methods are used to develop risk prediction models for use in bowel cancer screening. Standard statistical methodology in the form of logistic regression is initially used to integrate risk predictors with the FIT. This is then extended to a machine learning approach; an artificial neural network. The final approach considered is time to event analysis (survival analysis) where Cox Regression is implemented as the risk prediction model and further parametric models are explored. The 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis' (TRIPOD) guidelines for developing and reporting a multivariable prediction model are considered throughout. Where possible the recommendations and framework of the PROGRESS research group have also been implemented. Key themes of the research consider the cross-overs between risk prediction modelling and diagnostic accuracy research, specifically applying a diagnostic risk prediction model as a test and assessing the performance.

The studies in this thesis aim to contribute to determining the value of risk adjusted colorectal cancer screening using the FIT.

2.0 Outline of Thesis

Chapter two is a systematic review to understand whether risk prediction models that combine the FIT with other predictors performs better than regular FIT screening for colorectal cancer in terms of model performance and test accuracy. This systematic review crosses diagnostic accuracy and risk prediction model review boundaries and considers aspects and reporting guidelines from both. This crossover between risk prediction and diagnostic accuracy is carried forward as one of the main themes of the thesis.

Chapter three is a risk prediction model development study combining routinely available risk factors from the BCSS (the computer system used for the NHS BCSP in England) with the FIT result to assess whether model performance and test accuracy are improved. The modelling procedure applied in this chapter was logistic regression as a conventional approach in statistical methodology commonly used to build risk prediction models. One of the major advantages of this approach is the use of routine data requiring no additional data collection from participants, which improves data completeness and reduces participant burden. Further to this, the outputs of logistic regression are well understood and can be applied simply to an external dataset/screening population.

Chapter four then considers a machine learning approach to developing a risk prediction model in the form of a neural network to determine whether this performs better than standard statistical methods. This method does not require the strong assumption of linearity, as seen with logistic regression, and allows combinations of predictors to be combined without underlying knowledge of their interactions or relationship with the output variable. The same routine predictors are considered as those investigated in Chapter 3 and cross-validation is used to aid comparison. Individualised risk probabilities were produced for each patient and compared with the logistic regression risk adjusted FIT model and FIT only. The transparency of reporting this machine learning method is focussed upon by providing the full risk equation and detailing the model building process.

Chapter five investigates the use of an electronic primary care database to improve colorectal cancer screening referral decisions. Potential predictors from GP databases which may enhance a future risk adjusted model were investigated along with the data completeness of these predictors. Databases from primary care have a richer level of data

which may add a further dimension to a risk based prediction model to improve colorectal cancer screening. For instance, data are available on, symptoms, diagnoses, prescriptions, laboratory test results, lifestyle parameters and anthropometrics. The anonymised GP record database, 'The Health Improvement Network' (THIN) was used to extract data for the patients of interest and for model development. There are links between GP data and screening data, it is possible that further information could be drawn onto the screening system to contribute to decision making in screening. A further modelling approach is considered in this chapter in the form of survival analysis to predict the diagnosis of colorectal cancer within 2 years of the latest FOBT result. Using time to event analysis for longitudinal health records is a more efficient use of data than logistic regression since outcomes and exposures can occur at multiple time points for each participant. The semi-parametric Cox-Regression model was used to assess the association of >30 clinical features with colorectal cancer and polyps for an average risk screening population. A multivariable risk model was then developed using Cox-Regression and extended by considering the model fit of parametric models. The baseline survival was estimated for the Cox Regression model to give absolute risk probabilities for each individual.

Chapter six describes the methods used to extract valid data from THIN that was used for the analysis in Chapter 5. This included the development of a method to define acceptable periods of NHS BCSP notifications for practices receiving electronic results – the Acceptable Electronic BCSP (AEB) date. This AEB date was used to derive an average risk BCSP cohort for analysis from THIN, for data quality assurance, practice eligibility and to define the patient start dates for follow up in Chapter 5. The methods used to derive this date can be applied in future colorectal cancer screening studies, other cancer screening programmes or further electronic health record databases. In order to extract data for a research study from THIN for symptoms and diagnoses, Read code lists specifying the defined diagnosis/symptom need to be constructed. For prescriptions, drug code lists are also used to extract this information. Methods are presented for selected examples of diagnoses, symptoms and prescriptions. The additional health data file which houses data such as laboratory test results is more complex and a methods strategy needs to be constructed in order to extract the data of interest, in the units of interest and within the acceptable/relevant range. The methods for extracting haemoglobin concentration are presented as an example in this chapter.

All these studies are tied together for the thesis discussion (**Chapter Seven**) with a summary of the findings from each chapter and the corresponding future perspectives in this area of research.

3.0 References

1. Ahmad AS, Ormiston-Smith N, Sasieni PD. Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960. Br J Cancer. 2015;112(5):943-7.
2. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. 2012 [Available from: http://globocan.iarc.fr,.
3. Cancer Research UK. Lifetime Risk of Bowel Cancer 2013 [Available from: (http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-Four).
4. Cancer Research UK. Lifetime Risk of Cancer 2012 [cited 2014 8th November]. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats/incidence/risk/statistics-on-the-risk-of-developing-cancer.
5. National Screening Committee Public Health England. NHS population screening explained 2018 [Available from: https://www.gov.uk/guidance/nhs-population-screening-explained.
6. Raffle A, Gray M. Screening Evidence and Practice. New York, United States: Oxford University Press; 2007.
7. Wilson J, Jungner G. Principles and practice of screening for disease Geneva: World Health Organisation; 1968 [Available from: http://apps.who.int/iris/bitstream/handle/10665/37650/WHO_PHP_34.pdf?sequence=17.
8. Holland WW, Stewart S. Screening in Disease Prevention: What Works? 2005: Radcliffe Publishing Ltd; 2005.
9. National Screening Committee Public Health England. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme 2015 [Available from: https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme.
10. Hewitson P, Glasziou P, Irwig L, Towler B, Watson E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. The Cochrane database of systematic reviews. 2007(1):Cd001216.
11. UK National Screening Committee. Bowel Cancer: The UK NSC policy on Bowel Cancer screening in adults 2013 [Available from: http://www.screening.nhs.uk/bowelcancer.
12. UK National Screening Committee. Bowel cancer screening across the UK 2013 [Available from: http://www.screening.nhs.uk/bowelcancer-compare.
13. van Rossum LG, van Rijn AF, Laheij RJ, van Oijen MG, Fockens P, van Krieken HH, et al. Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. Gastroenterology. 2008;135(1):82-90.
14. Launois R, Le Moine JG, Uzzan B, Fiestas Navarrete LI, Benamouzig R. Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. European journal of gastroenterology & hepatology. 2014;26(9):978-89.
15. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. International journal of cancer Journal international du cancer. 2014;135(5):1203-12.
16. Digby J, Fraser CG, Carey FA, McDonald PJ, Strachan JA, Diament RH, et al. Faecal haemoglobin concentration is related to severity of colorectal neoplasia. Journal of clinical pathology. 2013;66(5):415-9.

17. Garcia M, Mila N, Binefa G, Benito L, Gonzalo N, Moreno V. Fecal hemoglobin concentration as a measure of risk to tailor colorectal cancer screening: are we there yet? European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2015;24(4):321-7.
18. Halloran SP, Launoy G, Zappa M. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition--Faecal occult blood testing. Endoscopy. 2012;44 Suppl 3:Se65-87.
19. Stracci F, Zorzi M, Grazzini G. Colorectal Cancer Screening: Tests, Strategies, and Perspectives. Frontiers in Public Health. 2014;2:210.
20. Carroll MR, Seaman HE, Halloran SP. Tests and investigations for colorectal cancer screening. Clinical biochemistry. 2014;47(10-11):921-39.
21. U. S. Preventive Services Task Force. Screening for colorectal cancer: Us preventive services task force recommendation statement. JAMA. 2016;315(23):2564-75.
22. Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J, et al. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. Gut. 2016.
23. Cooper JA, Moss SM, Smith S, Seaman HE, Taylor-Phillips S, Parsons N, et al. FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2016;18(7):650-3.
24. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014.
25. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ (Clinical research ed). 2013;346:e5793.
26. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
27. Omata F, Shintani A, Isozaki M, Masuda K, Fujita Y, Fukui T. Diagnostic performance of quantitative fecal immunochemical test and multivariate prediction model for colorectal neoplasms in asymptomatic individuals. European journal of gastroenterology & hepatology. 2011;23(11):1036-41.
28. Tao S, Haug U, Kuhn K, Brenner H. Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. British Journal of Cancer. 2012;106(8):1424-30.
29. Watson J, Shaw K, Macgregor M, Smith S, Halloran S, Patnick J, et al. Use of research questionnaires in the NHS Bowel Cancer Screening Programme in England: impact on screening uptake. Journal of medical screening. 2013;20(4):192-7.
30. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2):e1001381.
31. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ (Clinical research ed). 2009;338:b604.
32. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart (British Cardiac Society). 2012;98(9):691-8.
33. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS medicine. 2014;11(10):e1001744.

34. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). Journal of the National Cancer Institute. 2005;97(16):1180-4.
35. Cochrane Colloquium Vienna, editor PROBAST: a risk of bias tool for prediction modelling studies. Cochrane Colloquium; 2015; Vienna.
36. Moons KM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.
37. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ (Clinical research ed). 2013;346:e5595.
38. Thangaratinam S, Allotey J, Marlin N, Mol B, Von Dadelszen P, Ganzevoort W, et al. Development and validation of Prediction models for Risks of complications in Early-onset Pre-eclampsia (PREP): a prospective cohort study. Southampton (UK): NIHR Journals Library; 2017 Apr. 2017;Health Technology Assessment, No. 21.18.) Available from: https://www.ncbi.nlm.nih.gov/books/NBK425688/ doi: 10.3310/hta21180.
39. Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, et al. A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. Gastrointestinal Endoscopy. 2015;81(3):719-27.
40. Chen LS, Yen AM, Chiu SY, Liao CS, Chen HH. Baseline faecal occult blood concentration as a predictor of incident colorectal neoplasia: longitudinal follow-up of a Taiwanese population-based colorectal cancer screening cohort. The Lancet Oncology. 2011;12(6):551-8.
41. Digby J, Fraser CG, Carey FA, Diament RH, Balsitis M, Steele RJ. Faecal haemoglobin concentration is related to detection of advanced colorectal neoplasia in the next screening round. Journal of medical screening. 2017;24(2):62-8.

Systematic Review of Risk Prediction Models Combining the FIT Result for Colorectal Cancer Screening

Authors: Jennifer Cooper (JC), Karoline Freeman (KF), Rebecca Crosby (RC), Chris Stinton (CS), Nick Parsons (NP), Sian Taylor-Phillips (STP).

Information Specialist/Subject Librarian: Samantha Johnson (SJ)

PROSPERO Registration Number: CRD42016032526

https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=32526

ABSTRACT

Objectives

To systematically review studies that develop or validate risk prediction models which combine the faecal immunochemical test (FIT) result with other predictors for colorectal cancer screening referral decisions. Predicting the risk of current colorectal cancer using diagnostic risk prediction models allows screening programmes to determine which individuals are at highest risk for referral to colonoscopy. This approach could potentially improve cancer detection rates and allow more effective use of colonoscopy resources. The primary objective of the review was to assess model performance as well as test accuracy measures where applicable. Secondary objectives included identifying different statistical methodologies used to combine predictors in the models (logistic regression, survival analysis, machine learning algorithms), types of predictors included (demographic characteristics versus lifestyle factors, biomarkers or lab results), how the model would be applied in practice and how it is presented for use (equation, nomogram, risk score).

Design

The following bibliographic databases were searched using a combination of medical subject headings, key words and recommended search filters: MEDLINE, EMBASE, Web of Science and the Cochrane Library. Clinical trial databases were also searched to identify any relevant ongoing studies as well as databases of conference proceedings for any relevant unpublished studies. Reference lists of included studies and relevant reviews were searched and experts contacted for any additional papers. Two reviewers independently sifted titles and abstracts and subsequently full texts using predefined criteria. Data extraction was performed by one reviewer and checked by a second reviewer. Recently developed methodology and guidance for systematic reviews of risk prediction models

were followed. This included using a pilot version of the Prediction study Risk Of Bias Assessment Tool (PROBAST). Domains from QUADAS-2 were used for quality assessment if the model was applied as a test. The review was primarily concerned with identifying diagnostic risk prediction models that allow individualised risk predictions to be obtained. Models were considered for inclusion if they combined the FIT, allowed an individualised risk prediction and assessed model performance. Test accuracy measures were considered additionally if assessed. Primary outcomes were therefore model performance (calibration, discrimination) and where applicable test accuracy (sensitivity, specificity).

Results

The searches identified 5,671 articles for title and abstract sifting after removing duplicate references. 54 full text articles were then reviewed for inclusion using predefined criteria. Eight studies were included for data extraction and quality assessment. The heterogeneity in statistical methodology, predictors included in the models, populations and outcomes reported meant a meta-analysis of model performance and test accuracy was not possible and so a narrative synthesis was performed. Discrimination ranged from 0.676-0.960 for risk adjusted FIT (reported in 6/8 studies) and 0.683-0.902 for FIT only (reported in 4/8 studies). Calibration using the Hosmer-Lemeshow statistic ranged from 0.276-0.940 for risk adjusted FIT (reported in 3/8 studies) and calibration plots were presented in just one study. Where test accuracy measures were included (4/8 studies) sensitivity ranged from 21.9% to 88.0% for risk adjusted FIT at a range of set specificities from 90-97.7%. FIT-only sensitivity ranged from 19.7% to 82.0% at the same specificities. Additional metrics included net reclassification improvement (reported by 2 studies). Five out of eight studies used logistic regression, two further studies used a modified version of logistic regression and 1/8 used an accelerated failure time model (survival analysis). Half the included studies used some form of internal validation. Predictors used in the models varied from demographic characteristics, biomarkers and lab results as well as additional information obtained from questionnaires. The most consistently used predictors across the models were sex and age. PROBAST identified that all studies were at high risk of bias. For those studies which applied the model as a test (4/8), a high risk of bias was identified in 50% (8/16) of the domains using QUADAS-2.

Conclusions

Although it could not be tested formally due to the underlying heterogeneity, there is some evidence to suggest that including additional factors with the FIT result can improve model performance and test accuracy comparing FIT with risk adjusted FIT models. The models identified are at a development stage and need further assessment and validation before being applied in practice. Lab results and biomarkers tended to give higher discrimination values and test accuracy metrics but a significant improvement was also seen when using simple routinely available predictors such as age and sex which are more readily available. This suggests a reasonable improvement could be achieved using demographic factors alone without additional laboratory testing. Electronic Health Records and routine data could provide a convenient source of data for model development and validation. Future research should incorporate predictors which were consistent across these studies as well as predictors in risk prediction models for colorectal cancer without the FIT of which systematic reviews have been recently completed. Model development and reporting of the studies would benefit from using the TRIPOD guidelines as statistical analysis and predictor selection was often rated at high risk of bias using the PROBAST tool. Furthermore, external validation and impact studies are required before implementing any of the existing models. All the identified studies used standard statistical methodology to develop models; machine-learning algorithms have been shown to have similar or superior performance and could be assessed for performance alongside logistic regression to determine any improved outcomes.

1.0 BACKGROUND AND RATIONALE

1.1 Colorectal Cancer and Screening

Screening for colorectal cancer aims to detect the disease at earlier stages when it is easier to treat and cure. A Cochrane systematic review has shown that colorectal cancer screening using the faecal occult blood test (FOBT) can reduce the risk of mortality by 16%.[1] Countries worldwide including Australia, The Netherlands, Korea and the UK have implemented colorectal cancer screening at national, regional and local levels using either the guaiac faecal occult blood test (gFOBT), the faecal immunochemical test (FIT), optical colonoscopy or flexible sigmoidoscopy.[2] The most commonly used screening tool however is the FOBT.[3]

1.2 The FIT versus the Guaiac Based Test

Both the gFOBT and the FIT detect blood present in faeces as a marker for colorectal cancer and advanced adenomas. The more recently developed FIT has been shown to be analytically superior to the gFOBT and has been recommended for screening programmes worldwide by the 2010 European guidelines for quality assurance in CRC screening and diagnosis.[4] A systematic review and meta-analysis has shown that the FIT has a sensitivity of 87.2% and a specificity of 92.8% for colorectal cancer whereas the guaiac based test has a much lower sensitivity at 47.4% and a specificity of 92%.[5] The first random comparison between the guaiac and immunochemical based tests has also shown that the immunochemical test has a significantly higher advanced adenoma and cancer detection rate.[6]

Apart from the improved accuracy of the test, the FIT has several advantages over the guaiac based test including ease of use (with only one sample required instead of three), improved screening uptake and cancer detection rate,[6 7] is more specific for human haemoglobin and as such does not require dietary restriction. Perhaps most importantly, the test allows quantification of the result giving the amount/concentration of haemoglobin present per gram of faeces. This FIT concentration has shown to relate to risk and stage of colorectal cancer.[8 9]

1.3 Combining Risk Stratification with the FIT

The haemoglobin concentration provided by the FIT has been shown to be related to the risk of CRC, with a higher result indicating a greater risk. Yen *et al.*[9] have shown that the FIT haemoglobin concentration is an independent predictor for colorectal neoplasia risk. The concentration has also been related to severity of colorectal cancer lesions.[8 10]

Risk stratification using prediction models, risk scores and indexes to tailor screening in average risk individuals has been investigated.[11 12] It has been suggested that the FIT concentration can be included in individual risk assessment to improve the effectiveness of screening strategies.[8 10 13] Stegeman *et al.*[14] suggests that risk factors could be used to either target high-risk populations for screening through risk stratification or screening could be tailored by calculating risk for individuals. For instance, those who are at greater risk can commence screening at a younger age or be screened more regularly. The American College of Gastroenterology Guidelines based on a systematic review of the literature recommend CRC screening to start at age 45 for African-Americans compared with age 50 for other groups.[15] The guidelines also support screening heavy smokers and those at risk of obesity earlier.

Diagnostic risk prediction models which determine the probability of CRC can be used to improve the test accuracy during screening. There is some evidence to suggest that combining the FIT with individual risk factors enhances the accuracy of the test. For example, a study by Stegeman *et al.*[16] showed that a risk based model improved sensitivity from 32% to 40%. Risk factors for CRC can include any predictors which are associated with a higher or lower risk of CRC. Risk factors known to increase the likelihood of being diagnosed with CRC include, increasing age, male sex, lifestyle factors such as smoking, and increased alcohol consumption. Several different biomarkers (proteins, RNA and DNA) for CRC have also been identified along with single nucleotide polymorphisms (SNPs) from the Colorectal cancer GENeTics (COGENT) study which could be used in risk models.[17-19]

Why combine the FIT with other factors for this systematic review?

There are risk prediction models which have been developed and validated for colorectal cancer which do not combine the FIT result.[20 21] However, the discrimination power of models which do not include FIT are likely to be significantly lower. For example, Kaminski

et al.[21] develop and validate a logistic regression model which includes age, sex, family history of colorectal cancer, cigarette smoking and body mass index to predict the presence of advanced colorectal neoplasia. The model was well calibrated with moderate discriminatory power (AUC ROC 0.62).[21] The authors suggest combining predictors such as the ones identified in this model with FIT or blood based biomarkers to improve discrimination and referral selection. The model by Stegeman *et al.*[16] on the other hand which combines the FIT with calcium intake, family history and age had a higher discrimination with an AUC of 0.76.

FIT on its own, without other predictors may fail to detect intermittent bleeding or smaller lesions which may not bleed. Lab results such as including abnormal blood cell results with the FOBT have been shown to improve sensitivity for detecting colorectal cancer.[22] Additionally to this, the FOBT has been shown to be less sensitive in females,[23] and it has been suggested that sex specific cut off values for the FIT are used in screening porgrammes.[24 25] This suggests that lab test results as well as demographic factors may help to enhance the performance of the FIT. Individual risk prediction is usually low when only considering one factor and should integrate several parameters.[26]

1.4 Why is it important to do this review?

1.4.1 To build on previous models before developing a new one

Before developing a risk prediction model it is best practice to build upon what has already been developed previously.[27 28] For instance, there may be an internally validated model which could be suitable for the target population and the model could be externally validated and assessed in an impact study. Before implementing a model in practice it is important it has been externally validated and assessed in an impact study to ensure it improves decision-making and therefore patient outcomes. Alternatively, predictors which are found consistent in previous studies can be considered in future model development research.

There is currently an abundance of model development studies, fewer validation studies and very few impact studies.[29] A systematic review of predictive performance is required to determine the models predictive ability in different case mixes, settings, locations and to assess whether adjustments can be made to the model.[28] The guidance for systematic

reviews of therapeutic interventions is well developed with growing guidance for systematic reviews of diagnostic test accuracy. Prediction model systematic review methodology is still being developed with guidelines, checklists and quality appraisal tools being published over the last few years. More recently, methods for meta analyses have been considered for example with independent patient data.[30]

1.4.2 Improved accuracy of the test and cancer detection rates

Despite the relationship between FIT concentration and risk of CRC, screening programmes dichotomise the result and refer patients over a certain cutoff for colonoscopy and place those under the cutoff back into the screening pool. This wastes potential information relating to risk which could be used to personalise screening strategies to improve patient outcomes and screening performance. For instance, haemoglobin concentration has been found to be affected by participant demographics such as age and sex,[13 31-34] supporting tailored screening strategies. The study by Stegeman *et al.*[16] showed improved sensitivity using a risk based model over the FIT alone from 32% to 40% and an increase in cancer detection.

1.4.3 Colonoscopy Capacity

The FIT has recently been piloted in the UK in 40,000 participants,[7] and has been recommended to replace the guaiac test by the National Screening Committee (NSC) as of January 2016.[35] The results show that the uptake for the FIT was higher than the guaiac test (66.4% versus 59.3%) and the FIT detected more cancers (0.27% versus 0.12%) and advanced adenomas (1.73% versus 0.35%).[36] Due to an increase in the uptake of the test as well as increased test positivity (7.8% at a cutoff of 20µg/g versus 1.7% with the guaiac test), the FIT may put additional pressure on already limited colonoscopy capacity.[7] For instance, Ireland and the Netherlands have recently introduced the FIT for their screening programmes and had to alter their screening referral criteria due to an added strain on colonoscopy resources.[37 38] Future strategies for CRC screening will need to ensure positivity thresholds are set appropriately and screening referrals are recommended for those at greatest risk to ensure effective colonoscopy use. The NSC have recommended research into setting an appropriate threshold as colonoscopy capacity and uptake increases.[35] Further work from the UK pilot study will investigate different positivity thresholds and individualised thresholds according to patient characteristics.[36]

A risk-based approach to screening could offer several advantages including an increased detection of early stage cancers and their precursors as well as minimizing the number of false positives and negatives. In addition, by identifying people at higher risk instead of those at low risk, the use of available resources is optimised.[10 11 16]

1.5 Related Research and Systematic Reviews

Ma and Ladabaum[12] conducted a systematic review of risk prediction models, which could be used to personalise CRC screening. This systematic review looked at risk prediction models using clinical and demographic predictors. Some of the prediction models included sigmoidoscopy and colonoscopy results but not FOBT or FIT results. Usher-Smith *et al.*[39] carried out a systematic review for risk prediction models which predict the future risk of primary colorectal cancer for asymptomatic individuals. The current systematic review however proposes to investigate current undiagnosed colorectal cancer to assist with colonoscopy referral decision making. A systematic review on risk prediction models for colorectal cancer in people with symptoms has also been completed but this review focused on models which could be used for patients in primary care.[40] This population is treated differently to a screening population in terms of risk and the NICE guidelines for referral.

The accuracy of faecal immunochemical tests for CRC in asymptomatic, average risk adults has been investigated by Lee *et al*.[41] Subgroup analyses were performed on the number of FIT samples, cutoff value for a positive FIT test, FIT brand and reference standard. However, the review did not look at the effect of risk based testing or the use of risk algorithms on diagnostic accuracy. Finally, a systematic review has been carried out to identify biomarkers for early colorectal cancer detection and found that combinations of both fecal biomarkers and serum biomarkers led to higher test accuracy measures such as sensitivity and specificity.[42] This review however did not look specifically at risk prediction models or consider other predictors.[42]

An increasing number of studies have been published in this area in recent years. A Research Group in the Netherlands are investigating FIT based CRC screening and risk stratification. Over the last few years at the World Endoscopy Organization CRC Screening meeting, there has been increased research relating to risk based screening. In addition,

the UK FIT research team are investigating whether combining routinely available risk factors with the FIT improves test accuracy. The USA include in their guidelines risk based assessment by subgroups when considering what age to commence screening.[15] Ireland and Scotland have implemented FIT into their screening programmes. This method could also combat the problem of increased pressures on colonoscopy resources and optimise the benefits and harms of screening.

The proposed systematic review will be investigating studies which have combined risk predictors with the FIT result in a diagnostic risk prediction model and will pilot an early version of the recently developed prediction model risk of bias assessment tool (PROBAST) to investigate risk of bias and systematic review applicability.[43]

1.6 Objectives

Primary Objective

The primary objective was to perform a systematic review to identify risk prediction models which combine the FIT result for colorectal cancer screening referral decisions and to determine whether they perform better than regular screening using the FIT alone.

Further Objectives

- Identify all risk scoring systems (risk models, scores, clinical decision rules and other algorithms) which combine the FIT result for colorectal cancer screening.
- To determine the characteristics of predictive models (including which predictive factors are included in the models or used in the recall algorithm)
- To identify the important predictors for CRC detection in the prediction models
- To describe key methodologies used to combine risk factors into the screening recall algorithm (e.g. logistic regression, survival analysis, machine learning approaches and clinical decision rules etc.)
- To describe how the models are presented for use (equation, nomogram, risk calculator etc.)
- To assess the model performance of diagnostic predictive models (including the performance measures: calibration, discrimination and classification)

- To assess test accuracy of the predictive models (including sensitivity and specificity at particular cutoffs and the area under the curve of the Receiver Operating Characteristic Curve)
- To determine whether risk prediction models combining the FIT have better test accuracy than screening using the FIT alone.
- If the data permits, to perform a meta-analysis of test accuracy (sensitivity and specificity) and model performance (discrimination and calibration)
- To make recommendations on the use of risk prediction models in screening for CRC

2.0 METHODS

This review considered risk prediction model studies which combined the FIT result with at least one other predictor to produce individualised risk predictions for colorectal cancer/adenomas for diagnostic testing referral decisions. The risk models needed to give an individualised risk prediction and ideally a measure of absolute risk.

Analysis focused upon risk prediction model performance and was supplemented with test accuracy measures. Risk of bias in studies was assessed using criteria from both PROBAST[43] (prediction modelling studies) and QUADAS-2[44] (in those assessing a test accuracy component). A similar approach was undertaken in a previous HTA report where criteria relating to prognostic studies as well as diagnostic accuracy studies was used to determine the overall quality of the studies.[45]

Systematic reviews of prognostic and diagnostic risk prediction models is a new and evolving area and methods are being developed for these types of reviews. Exemplar reviews have been pursued for prediction model reviews including the protocol by Pace *et al.*[46] The Cochrane Prognosis Methods Group established in 2008 are developing methodological tools and resources in this area for use in systematic reviews of prediction modeling studies.[47] Several of these tools were used or piloted within this systematic review including the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies (CHARMS checklist) which was used to form the review question for the systematic review as well as for critical appraisal of prediction studies and data extraction.[27] In addition, an early version of PROBAST was piloted to assess the risk of bias of prediction modeling studies and the applicability of the study to

the systematic review.[43] The full PROBAST assessment tool and the explanation and elaboration document will be published in 2018. PRISMA guidelines will be followed to report the systematic review.[48]

2.1 Criteria for considering studies for this review

The CHARMS checklist was used to guide the review aim, search strategy and inclusion and exclusion of studies used in the review across seven different domains (See **Table 1**).[27] Any risk prediction model which has combined the FIT result with at least one other predictor to determine the risk of CRC was included in this review. For instance, a study may have used an individual risk questionnaire and separately combined this information with the FIT result to determine who to refer for colonoscopy.

Item	Comment
1. Prognostic versus diagnostic prediction model	Diagnostic prediction model. The aim is to review models to predict current disease status for CRC screening.
2. Intended scope of the review	The scope of the review and intended purpose of the models reviewed in it. Risk prediction models combining the FIT result to inform referral to diagnostic testing
3. Type of prediction modeling studies	**1a** – Prediction model development (ideally with internal validation) **1b** – Prediction model external validation (with possible updating) **1c** - Prediction model development and external validation **1d** – Developing/validating a model and then applying as a test **1e** – Applying the risk prediction model as a test (impact study)
4. Target population to whom the prediction model applies	The target population relevant to the review scope. Average risk screening population or representative of a screening population Both men and women aged 40-75 (10% outside screening range is acceptable) Both organised screening and opportunistic screening.
5. Outcome to be predicted	The outcome of interest to be predicted. Specific diagnostic target disease – colorectal neoplasia (CRC and advanced adenomas) detected at colonoscopy. The model should predict; colorectal cancer, adenoma and polyp diagnoses.
6. Time span of prediction	Predicting current disease status in screening for referral to colonoscopy
7. Intended moment of using the model	The systematic review may focus on models to be used at a specific moment in time. - Model to be used around the time of FIT screening to identify high risk individuals for further diagnostic testing such as colonoscopy

Table 1: Key items to guide the framing of the review aim, search strategy, and study inclusion and exclusion criteria based on the CHARMS checklist. CHARMS: CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modeling Studies.

2.1.1 Selection and Inclusion Criteria

The selection criteria was based on PICOTS (Participants, Intervention, Comparator, Outcome, Timing, Setting) which is recommended for use in risk prediction modelling studies.[27 28] PICOTS is an amended version of PICO which includes a temporal element for the moment of using the model.

Participants

Average risk screening population or representative of a screening population Both men and women aged 40-75 (10% outside screening range is acceptable). Screening programmes worldwide target individuals within this age range.[49] Both organised and opportunistic screening will be considered since screening is implemented in different ways across countries.

Intervention: Index test combined within a risk prediction model:

Where the FIT has been combined with other risk factors/predictors in a risk prediction model to aid referral decisions to diagnostic testing.

All risk factors/predictors will be investigated including demographic, clinical and lifestyle factors. For example, predictors could include age, sex, socioeconomic deprivation, dietary factors, physical activity, alcohol consumption, smoking status, previous screening history, medication use, BMI, hypertension etc. Genetic markers and biomarkers including DNA proteins, messenger RNA and microRNA will also be included.

Comparator Tests: Index Test alone

Studies may or may not have used FIT as a comparator. FIT alone can be compared either applied as a test or applied within a model. FITs of all brands both qualitative (positive/negative) and quantitative (provides a concentration of haemoglobin) for the detection of CRC and advanced adenomas (AAs). Qualitative FITs are generally used as point of care (POC) tests and use a lateral flow immunochromatographic analysis technique.[2] This type of FIT relies on visual interpretation with the cutoff set by the manufacturer. Quantitative FITs on the other hand use an immunoturbidimetry technique for analysis.[2] This test provides a continuous result and the cutoff can be set to fit local contexts taking into account prevalence of disease and colonoscopy resources.

Outcomes (Target condition/reference standard/performance measures)

Models to predict presence of colorectal cancer, adenomas or polyps. Colorectal advanced neoplasia includes both CRC and advanced adenomas. Advanced adenomas have a high risk of developing into cancer and are defined by the European Guidelines for Quality Assurance in CRC and Screening Diagnosis (2010) as one which is 10mm or over, or contains high grade mucosal neoplasia or contains a villous component.[50]

Where diagnostic accuracy parameters are included, the reference standard can be either:
a) Colonoscopy for the detection of CRC or advanced adenomas
b) At least two-year longitudinal follow-up using clinical records (e.g. cancer registries, GP records etc.

The primary outcomes for the review will be model performance and test accuracy parameters in relation to CRC detection at colonoscopy. For instance, discrimination, calibration and classification will be identified for model performance whereas sensitivity, specificity, AUC ROC, positive prediction value (PPV) and negative prediction value (NPV) will be identified for test accuracy.

Timing

As a diagnostic prediction model to predict the probability of current disease before a reference (gold) standard test that has not yet been performed.

Setting/Role

Intended role of the model is to determine risk for referral to diagnostic testing or colonoscopy in a bowel cancer screening/average risk population setting.

2.1.2 Types of studies included in the review

This review will include studies of any design that develop or validate a risk prediction model which combines FIT with other predictors to predict risk of colorectal cancer/adenoma diagnosis. Risk prediction modeling studies include: model development (ideally with internal validation), model external validation and impact studies. Some studies can develop a model and validate it within the same paper. However, the gold standard of model development and external validation would usually require a separate

dataset to validate and ideally separate research team as there is evidence to suggest model performance measures are inflated otherwise.[51] The inclusion and exclusion criteria for the abstract and title sift as well as the full text review are shown in **Table 2** and **Table 3** respectively.

The following categories will be considered in this review:

1a - Model development (ideally with internal validation)

1b - Model development with external validation

1c - External Validation

1d - Developing/validating a model and then applying as a test

1e - Applying the risk prediction model as a test (impact study incl. diagnostic accuracy).

2.1.3 Inclusion and Exclusion Criteria for abstract and title sift

Inclusion Criteria
Screening Test: The risk prediction model includes the FIT result and at least one other predictor
Model: The risk prediction model can be used to give individualised risk predictions ideally absolute measures of risk but relative levels of risk will also be included (e.g. linear predictor or scoring systems)
Outcome: The outcome of the prediction model is colorectal cancer, adenomas or polyps
Performance Measures: The study reports model performance parameters (calibration, discrimination and classification) or test accuracy parameters (sensitivity, specificity, PPV, NPV or area under the ROC curve)
Population: The risk prediction model is developed or applied on men and women aged between 40-75 representing the average risk screening population
Exclusion Criteria
Articles in languages other than English
Case Studies
Articles published before 1978
Non-human studies

Table 2: Inclusion and exclusion criteria for abstract and title sift

2.1.4 Inclusion and Exclusion Criteria for Full Text Sort

Study Design	**Y/N/U**
Studies of any design that develop, validate, update a diagnostic risk prediction model for colorectal cancer which combines FIT with at least one other predictor. Risk prediction modeling studies for this review include the following categories: **1a -** Model development (ideally with internal validation) **1b -** Model development with external validation **1c -** External Validation **1d -** Developing/validating a model and then applying as a test **1e -** Applying the risk prediction model as a test (impact study incl. diagnostic accuracy). *Note*: Studies may develop a model and externally validate it within the same paper.	
Model	
Definition of a diagnostic model in this review: Combination of FIT with at least one other predictor to predict individualised risk of colorectal cancer/advanced adenomas using a statistical model (can include neural networks, logistic regression, survival analysis/Cox Regression or other approaches i.e. a multivariable diagnostic study only)	
Intended moment of using the model: Can the model be used around the time of FIT screening to identify high-risk individuals for referral (further diagnostic testing such as colonoscopy)? i.e. before diagnostic testing	
Does the model assess more than just the association of predictors with outcome? (I.e. Exclude if it is only a logistic regression or Cox regression which is not applied/developed as a risk prediction model, it just presents ORs and HRs)	
Has the index test (FIT) been combined with at least one other predictor/risk factor* in a diagnostic risk prediction model? *Predictors could include; age, sex, socioeconomic deprivation, other test results, dietary factors, physical activity, alcohol consumption, symptoms, previous screening history, medication use, genetic markers, biomarkers, blood results etc	
Does the study produce a risk prediction model/scoring system/algorithm using statistical methods such as neural networks, logistic regression, survival analysis (Cox Regression), machine learning or other similar approaches?	
Screening Test (Index Test)	
The test assessed is the Faecal Immunochemical test (FIT) which includes all brands both quantitative (provides a concentration of haemoglobin) and qualitative (positive/negative). Examples of FIT brands include; OC-Micro/Sensor, OC-Light, Hemeselect, Flexsure OBT, FOB Gold, OC Hemodia, FECA-EIA, HM-Jack, Instant-View, Occultech, Ridascreen. *Note*: If the test is a guaiac based test, exclude. Brand names include: Haemoccult, Haemoccult II, Hemoccult Sensa, Fecatwin, Fecatest.	
Reference Standard	
Where diagnostic accuracy parameters are included (including impact studies) is the reference standard either: -Colonoscopy (small numbers of other diagnostic procedures acceptable as some individuals may not be suitable for colonoscopy e.g. CT/flexible sigmoidoscopy) -At least two year longitudinal follow up using clinical records (e.g. cancer registries, GP records etc)?	

Population	
Both men and women aged 40-75 years *representative* of an average risk screening population (mean/average age needs to be over 40 years so participants over 18 can be included/10% outside screening range is acceptable) *Note*: Symptomatic patients may be included if considered a part of a general screening population sample with symptoms used as predictors (Symptoms requiring further investigation suggests the population is primary care GP referral two week wait criteria. NICE guidelines list these symptoms.)	
Outcome (Target Condition)	
Does the model predict the following outcomes? -Advanced neoplasia which can include both CRC and advanced adenomas -Colorectal polyps	
Performance Measures	
Does the study include model performance parameters, test accuracy parameters, or both? a) Model performance parameters (calibration, discrimination (also called AUC ROC or c-statistic), re-classification) b) Test accuracy parameters in relation to CRC detection (sensitivity, specificity, 2x2 data, AUC ROC) *Note*: If analytical sensitivity and specificity has been reported (differs to clinical sensitivity), the population may not be appropriate for the review (these studies tend to just look at positive samples/random samples)	

Table 3: Inclusion and Exclusion criteria for full text assessment. If all criteria are 'Y' include the study, if there are any 'N's exclude the study, if there are any 'U's then discuss the study.

2.2 Search methods for the identification of studies

A search strategy was developed with the input of an information specialist (SJ) to ensure all relevant articles were retrieved from the appropriate databases. As recommended by the Cochrane handbook and recent guide to systematic review and meta-analysis of prediction model performance,[28] the search strategy incorporates the search filters suggested by Geering *et al.*[52] to identify diagnostic prediction studies in Medline. This strategy includes using the Ingui filter[53] in combination with the additional search string developed by Geering *et al.*[54] (using the Boolean operator 'OR'). This was combined with the disease of interest (CRC) and terms for the screening test. Further search terms which encompassed diagnostic accuracy were also included. See **Appendix 1** for the search strategies.

The search strategy was developed to maximize sensitivity (number of relevant studies identified over the total number of relevant reports on the topic) rather than focusing on

precision (number of relevant records identified by a search over the total number of records) as recommended by the Cochrane Handbook for DTA Reviews.[52]

Scoping Searches

Scoping searches were undertaken to determine the types of studies and the volume of studies relating to the research aim. Medline, Embase, and Web of Science were used for scoping purposes, to formulate a draft search strategy and inclusion/exclusion criteria.

Electronic searches

1. The electronic databases that were searched to identify published studies:

MEDLINE (via Ovid), EMBASE (via OVID), Cochrane Library (Wiley) (including the Cochrane Database of Systematic Reviews, the database of abstracts of reviews of effects (DARE) and the Health Technology Assessment database) and Thomson Reuters Web of Science using a combination of medical subject headings and key words.

The review was restricted to English language and limited to the date from which the first FIT was produced in 1978.[55]

2. To identify any relevant ongoing studies, the following trial registers were searched:
Clinical trial databases searched will include; Cochrane Central Register of Controlled Trials (CENTRAL), ClinicalTrials.gov, UKCRN Portfolio Database; WHO International Clinical Trials Registry Platform, ISRCTN.

Searching other resources

Grey Literature

Conference abstracts have been shown to overestimate the accuracy of a test.[56] However, since this is a niche field and all the current available evidence is desired for this review, authors of relevant conference proceedings were contacted for results or to obtain further information. The Zetoc database (The British Library), Science Citation Index and Conference Proceedings (Web of Science) were searched for relevant unpublished studies.

Reference lists

All reference lists of included papers and relevant reviews were searched and any appropriate papers considered for inclusion.

Correspondence

Experts were contacted to identify any further papers for review.

Sources of Heterogeneity

Brand of FIT

Qualitative (positive/negative) versus Quantitative (haemoglobin concentration FIT)

Cutoff value selected

Patient characteristics

Reference standard

Number of samples

Risk factors considered in the model

Predictive model methodology /statistical methodology

How the model is applied/presented (e.g. nomogram, risk groups, risk calculators)

2.3 Data collection and analysis

2.3.1 Selection of studies

After running searches on the above databases all citations were saved to Endnote X7 Software. Duplicate references were removed at this stage. Two reviewers independently screened the titles and abstracts (JC, RC) retrieved from the search against the inclusion and exclusion criteria. The full texts of those studies included were then also screened independently by two reviewers (JC, KF) and any discrepancies resolved with a third reviewer (NP, CS). Reasons for exclusion were recorded. The PRISMA study flow diagram depicting the screening process and reasons for exclusion are provided in the results section.

2.3.2 Data extraction and management

Data on 'study characteristics', 'model characteristics' and 'screening test' characteristics were extracted using a pre-specified data extraction table. The data extraction table was constructed using the CHARMS checklist,[27] previous systematic reviews of prediction models including Cochrane Exemplar Reviews,[45] the Cochrane Diagnostic Accuracy Handbook, the STARD statement [57] and QUADAS-2[44]. One reviewer extracted the data from the included studies and a second reviewer checked this information.

Data extraction included the following domains (the full data extraction form can be found in **Appendix 2**):

- Study characteristics (e.g. study design, country, setting)
- Source of data (e.g. cohort, case control, prospective, retrospective etc)
- Participants (e.g. inclusion and exclusion criteria, recruitment method and participant description)
- Outcomes to be predicted (e.g. definition and measurement of outcome)
- Candidate Predictors (E.g. number and type of predictors, definition of predictors
- Sample Size and Missing Data
- Model Development (e.g. modeling method (logistic regression, survival analysis, machine learning approaches) methods for selecting predictors)
- Model Performance (e.g. discrimination, calibration and classification measures)
- Model Evaluation (e.g. internal and external validation)
- Risk Stratification (How has risk been combined with the FIT and how has risk stratification been presented; as a score, as a probability etc)
- Diagnostic Accuracy Considerations (target disease definition, type and brand of FIT, reference standard)
- Test accuracy (diagnostic accuracy measures such as sensitivity, specificity, area under the receiver operating characteristic curve, cutpoint used)
- Results (presentation of the final multivariable model)
- Interpretation and discussion (Clinical applicability of the model, strengths, limitations etc)

2.3.3 Assessment of risk of bias

This systematic review was predominately a risk prediction model review and therefore PROBAST was used to assess methodological quality of the risk prediction model development and validation studies. Diagnostic accuracy measures if assessed within the same paper were considered an add on measure. Where the study considered a test accuracy component, QUADAS-2 was used to assess the methodological quality. The QUADAS-2 tool was tailored to this review using the FITTER standards[58] as guidance (**Appendix 4**). The Cochrane Risk of Bias Tool has been recommended by experts to use for model impact studies (comparative or intervention studies for different risk assessment).

An early pilot version of PROBAST[43] was used to assess risk of bias and the applicability of prediction modeling studies to the systematic review (Mallett *et al*, personal communication). This tool is based and developed from previous tools including the Quality in Prognostic Studies (QUIPS) tool,[59] and QUADAS-2[44] as well as from REporting recommendations for tumour MARKer prognostic studies (REMARK)[60] and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).[61] PROBAST was used independently by two reviewers to assess the quality of the diagnostic prediction modeling studies. PROBAST has five key domains which are assessed for risk of bias; participation selection, predictors, outcome, sample size and participant flow and analysis. The first three domains are also judged for applicability to the systematic review question. QUADAS-2 comprises 4 domains which are assessed in terms of risk of bias and include; patient selection, index test, reference standard and flow and timing. Applicability is also assessed using the first 3 domains.[44]

Appendix 4 includes an example of the tailored QUADAS-2 tool and key details for the pilot version of PROBAST.

2.4 Data synthesis

This review aimed to summarise the evidence of risk prediction models which combine the FIT for CRC screening. A narrative synthesis of the evidence was given using the data extracted and quality appraisal tools. Results and characteristics of individual studies were

presented in tabular displays for study characteristics, model characteristics and screening test characteristics in summary of results tables.

Study characteristics included population and setting, outcome measure of the model and sample size used for model development.

Model Performance

Model characteristics included type of predictors used in the model, statistical modeling approach, model performance parameters and how the model is presented/applied in practice which were summarized in a summary of results table.

If the data permitted and studies evaluated the same prediction model, a meta-analysis of calibration and discrimination was planned using DerSimonian and Laird's random effects meta-analysis.[62] A random effects model would have been required based on the probable methodological and clinical variability of the studies. Statistics which could be pooled to measure discrimination included the concordance (c) index, area under the curve and discrimination slope.[63] For calibration statistics, the O:E statistic, calibration-in-the-large and calibration slope can be pooled.[63] It is not recommended to pool results from the Hosmer-Lemeshow test and from the maximum likelihood as these measures depend on sample size.[63]

Test Accuracy

The screening test characteristics summary of results table included sensitivity, specificity, AUC ROC and cut-point used for the test and whether FIT only was used as a comparator. Cutoffs of the tests were converted to micrograms of haemoglobin per gram of faeces (μg/g) as recommended by the World Endoscopy Organization.[64] Test accuracy components were considered by four studies. Where risk based FIT was compared to using the FIT alone, linked ROC plots were produced where sensitivity and specificity are joined by a line in ROC space to show the relative test accuracy in each study. A linked forest plot of sensitivity and specificity without summary statistics was presented for only one study where it was possible to derive the 2 by 2 data (true positives (TP), true negatives (TN), false positives (FP), false negatives (FN)).

If the same prediction models were evaluated as tests, a meta-analysis was planned to combine estimates of test accuracy including sensitivity and specificity. Due to the heterogeneity of studies included in the review in terms of predictors included in the models, outcomes, statistical modelling procedure and underlying populations, the data did not permit a meta-analysis.

If a common threshold value (predicted probability) was used, results could be used to generate the summary sensitivity and specificity of the tests at that threshold. If a similar threshold was used in the identified studies then a bivariate model could be used for meta-analysis to give an average operating point across studies.[65] However, it was more likely for the threshold value to vary from study to study and therefore an estimation of a summary ROC curve would have been more appropriate by fitting a hierarchical SROC model.[66] By plotting the ROC curve, the accuracy of the test can be compared at a range of different thresholds. This would require the same risk prediction model to be investigated in each study since this model assumes the underlying ROC curve for each study has the same shape but if different predictors are used this would not be the case.

3.0 RESULTS

3.1 Search Results

The search produced 8,828 records of which 3,157 were duplicate records identified from Endnote and manually. This left 5,671 records to fully screen against the abstract and title inclusion/exclusion criteria. 54 articles were subjected to a full screen using the full inclusion/exclusion criteria (this included, 37 full text articles and 17 conference abstracts). 46 articles were excluded based on the following criteria; statistical model (n=19), screening test (n=8), population (n=5), study design (n=5). For abstracts, 8 studies had a full text article associated and 1 study the author could not provide additional information as the results were pending publication. Reasons for exclusion for each study are listed in **Appendix 3** and the PRISMA study flow chart is shown in **Figure 1.**

Studies identified by experts (Professor Tom Marshall and Professor Stephen Halloran) were also assessed against the pre-defined criteria for the review.[67-69] Boursi *et al.*[67] combined demographic, behavioural and past screening colonoscopy information but did not include the FIT. The other study by Kinar *et al*.[69] also combined lab based results with

other factors in a decision tree based prediction algorithm but included the guaiac based test as opposed to the FIT. Finally, the study by Birks *et al.*[68] investigated validating the previous model in a UK setting for primary care.

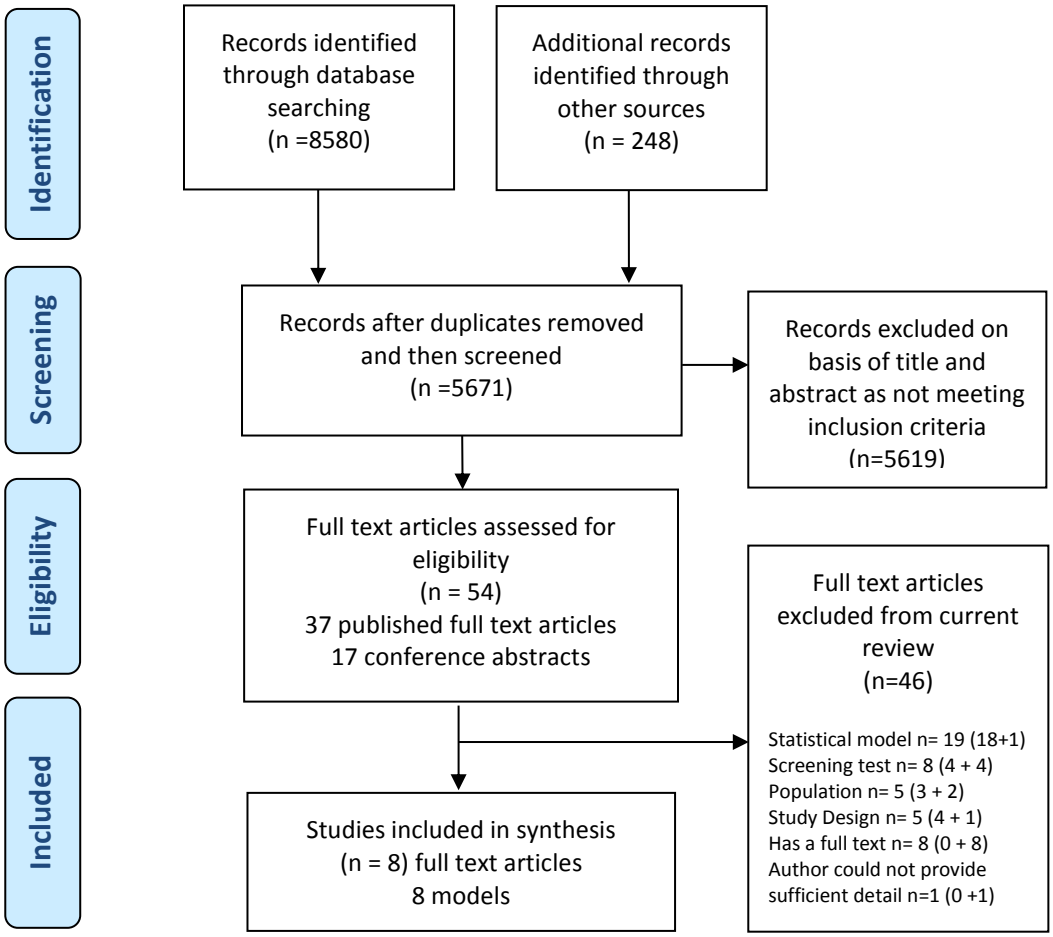


Figure 1: PRISMA Study Flowchart

3.2 Summary of included Full Text Articles

Stegeman *et al.*[16] investigated combining the FIT result with a risk questionnaire in CRC screening. The data were collected from a randomized screening pilot for using colonoscopy as a primary screening tool in the Netherlands. Risk factors selected to be included in the questionnaire came from previous analysis of variables for advanced neoplasia.[14] The final population size for model development was 1112 participants who completed both a risk questionnaire and the FIT. The statistical modelling strategy used logistic regression with backwards elimination.

Yen *et al.*[9] combined the FIT with conventional risk factors obtained from a questionnaire along with lab results (triglyceride levels). This study used a cohort from Taiwan invited to population based screening for colorectal neoplasia (n=54,921) to develop the model and another two datasets combined for external geographical validation (n=17,085) (two community based integrated screening programmes). This study produced nine models in total; one which was conventional risk factors only, another which was FIT adjusted for age and a combined model of FIT plus conventional risk factors. Each of these models was applied for the three different outcomes investigated: colorectal neoplasia, colorectal cancer and colorectal adenoma. Risk equations were presented for each model.

Omata *et al*.[70] combined the FIT result with routine predictors obtained from a general health checkup (age, sex, BMI). Data were collected as part of a cross sectional study investigating the optimal cut point of FIT in asymptomatic participants in Japan. The final sample size was 1085 (70% males) who completed both a FIT and colonoscopy. The modelling strategy was non-linear ordinal logistic regression using a three-category status of colorectal neoplasms as the dependent variable. The first objective of the study was to decide on the optimal cut-off point of the FIT – for this they selected an ideal cut-off value of FIT based on a value that maximised the sum of both sensitivity and specificity and clinical utility. This approach engineers good performance and not clinical reality.

Auge et al.[71] combined demographic characteristics (age and sex) with the OC-Sensor FIT. The population was FIT positive asymptomatic men and women from the first round of the Barcelona screening Programme (n=3109). A low threshold (20 µg/g) was used for positivity which gave a sample population at slightly higher than average risk. Logistic regression is used for model development and to delineate 16 risk categories and subsequently 3 risk levels based on the positive predictive values from the models.

Kim et al.[72] integrates the FIT with fecal Calgranulin B (CALB) and age using logistic regression. This study uses cases and controls from a previous study based in Korea where Western blot analysis of CALB was performed. Controls were recruited from participants at the Health Screening Program at the National Cancer Centre, Korea. Cases and controls from another independent patient cohort were also used. The modelling strategy uses a development set (81 cases, 51 controls), validation set (94 cases, 100 controls) and combined set (132 cases, 151 controls) to produce the overall risk prediction model.

Tao et al.[73] investigate combining blood based inflammatory markers with the FIT in a logistic regression algorithm to compare performance using the area under the curve. Data comes from participants recruited from an ongoing prospective screening study in Germany where screening colonoscopy is offered, as well as patients from the sub-study DACHS + where CRC patients are referred by GPs/gastroenterologists for surgery (enriched sample/case-control). The sample size used to produce the model includes 467 participants with available blood and stool samples. The study uses logistic regression as a tool to combine the blood based inflammatory markers and is mainly concerned with diagnostic accuracy over developing a risk prediction model but does provide individualised prediction.

Wietan *et al.*[74] combines the FIT with demographic factors (age, gender) using data from an organised screening programme in the Netherlands, Rotterdam-Rijnmond. The sample population used those with a positive test and successful colonoscopy (n=481). Data is combined in a logistic regression analysis to give a predicted probability of advanced neoplasia. Predicted probabilities are depicted in a figure with age as a continuous variable. The primary aim of this study was to assess how increasing the cutoff concentration for FIT and screening age subsequently affects colonoscopy yield, missed lesions and demand.

Karl *et al.*[75] investigated novel biomarkers to improve sensitivity of the detection of CRC using stool samples. The RIDASCREEN hemoglobin-haptoglobin FIT is combined with biomarkers TIMP-1 and S100A12 using Bayes logistic regression (BLR) as a mathematical model. The RIDASCREEN test was one of the FITs considered in recent NICE guidance.[76] Stool samples were obtained from two European multicenter studies in Germany; one of which was a preventative screening study at gastroenterology units and the other was used to recruit additional CRC patients from surgery units. The sample population comprised 252 controls, 113 advanced adenomas, 186 colorectal cancers. The main aim of the study was to examine the clinical performance of fecal S100A12 and other biomarkers. The study is therefore mainly concerned with diagnostic accuracy measures and uses BLR as a method to combine the biomarker and FIT results but does use posterior case probabilities to define a threshold for positivity (according to a specificity level of 95% or 98%).

3.3 Populations and Study Design

Models were developed using data collected from The Netherlands,[16 74] Japan,[70] Taiwan,[9] Spain,[71] Korea,[72] and Germany[73 75]. Study settings included data from population based colorectal cancer screening programmes, general health checkups and community based integrated screening. Data from case controls included a proportion of data from screening based settings. Included studies used a variety of study designs including prospective/retrospective cohort studies, cross sectional studies, case controls and enriched samples. **Table 4** summarises the study characteristics for the included papers.

The outcomes for the final models developed varied in spectrum across the studies from advanced adenoma[73] to colorectal cancer[72] and some considered a combined endpoint (e.g. advanced adenomas plus colorectal cancer as colorectal neoplasia)[9 16 70 71 74]. Combined endpoints are often included in screening studies due to the clinical benefit of detecting both early stage and later stage disease. Omata *et al.*[70] for instance considered significant neoplasia and the combined endpoint significant neoplasia or adenomatous polyps within the final Nomogram. Tao *et al.*[73] present model performance results for advanced adenoma only, since adding biomarkers to a model with FIT alone made no improvement to the detection of colorectal cancer. Karl *et al.*[75] present cross-validated results for colorectal cancer but also consider advanced adenomas and colorectal cancer separately when assessing a model combining both development and validation sets.

The TRIPOD classification of risk prediction model studies[61] was used as well as criteria developed for this review specifically (**Table 4**). Two of the studies are categorized as TRIPOD Classification 1b which is development and validation using resampling[16 70] although the methods for Omata *et al.*[70] are unclear. Two other studies use an independent dataset but for one of these studies[72] the final model presented combines both development and validation sets and for the other study the validation set uses new cases but the same controls from the development set[75]. These studies therefore cannot be considered a pure external validation study and are classified as a nonrandom split-sample development and validation study (TRIPOD Classification: 2b). Furthermore, these two studies[72 75] present cross validated performance results as well as performance determined externally but the final model presented for Kim *et al.*[72] is not internally validated. One of the studies was a development and validation study using separate data (TRIPOD

Classification: 3).[9] The final three studies are classified as development only (TRIPOD Classification: 2a) as internal validation is not reported.[71 73 74]

Four studies also consider applying the risk prediction model as a test and report test accuracy measures sensitivity and specificity.[16 72 73 75] These studies use FIT only as a comparator.

Study	Model of interest	TRIPOD Classification	Study Design	Country	Setting	Participant Description	Outcome	Sample Size for Model	Events
Stegeman, 2014	1d OC Sensor result + lifestyle questionnaire predictors	1b	Data from colonoscopy arm of a randomized screening trial – prospective cohort	The Netherlands	Population based colorectal cancer screening programme	Asymptomatic men and women 50-75 years old. Mean age 60.6 (SD 6.2) 543/1112 (49%) female	Presence of advanced neoplasia during colonoscopy	1112 underwent colonoscopy, completed a questionnaire and completed a FIT	101 (94 advanced adenoma and 7 cancer)
Yen, 2014	1b	3	Cohort invited to population screening for colorectal neoplasia using FIT – longitudinal follow up study	Taiwan. Keelung City for development dataset. Changhua and Tainan for validation dataset	Community based integrated screening	Subjects aged 40 years or older.	Risk for developing colorectal neoplasia (colorectal cancer and colorectal adenomas)	Development: 54,921 Validation: 883 colorectal cancers, 2028 adenoma cases and 14,174 disease free subjects	Development: 824 colorectal adenoma, 323 colorectal cancer Validation: 883 colorectal cancers, 2028 adenomas
Omata, 2011	1a	1b	Cross-sectional study of asymptomatic Japanese individuals undergoing a general health check-up	Tokyo, Japan	General health checkup with a proportion of patients undergoing colonoscopy	Asymptomatic patients from the general population. Mean age 64 years, 756/1085 (70%) male.	Three category status of Colorectal Neoplasms (CRN) – (A) No CRN, (B) AP excluding AAP (C) SN (significant neoplasia) AP (colorectal adenomatous polyps) AAP (advanced colorectal adenomatous polyps) SN includes AAP and CRC (colorectal cancer)	1085 completed full colonoscopy and QTFIT	393 cases of AP including 69 cases of AAP and 8 cases of CRC.
Auge, 2014	1a	1a	Retrospective study of a series of participants to the Barcelona colorectal cancer screening programme	Barcelona, Spain	First round of Barcelona colorectal cancer screening round	Asymptomatic men and women (50-69 years) participating in the Barcelona screening programme Median age – 60 43% women (1334/3109) 57% men (1775/3109)	Advanced colorectal neoplasia (colorectal cancer or high risk adenoma) versus non advanced colorectal neoplasia	3109 FIT positive participants undergoing colonoscopy	1147 high risk adenoma 294 colorectal cancer
Wieten, 2016	1a	1a	Data from an organised CRC screening programme in the Netherlands	The Netherlands, Rotterdam-Rijnmond	First round of CRC Screening Programme using the FIT	Average risk screening population aged 50-74 Median age 61 years, 48% of participants were male.	Advanced neoplasia (defined as CRC and advanced adenomas)	481 (positive test and successful colonoscopy)	164 advanced adenoma 29 CRC Advanced neoplasia = 193
Tao, 2012	1d	1a	Data from participants recruited from the BliTz study and satellite sub-study DACHS + Enriched sample/case control	Germany	The BliTz study is an ongoing prospective screening study where screening colonoscopy is offered. The DACHS study is an ongoing case-control study focusing on the role of colonoscopy in CRC prevention.	Men and women aged 55 and over in Germany from BliTz study. Participants from DACH study were CRC patients referred by GPs/gastroenterologists for surgery. These two studies were used to derive three groups of participants; those with no neoplasm, advanced adenoma, CRC.	Advanced adenoma OR colorectal cancer No colorectal neoplasm (44.4% male, mean age 61.9). Advanced adenoma (64.3% male, mean age 65.0). Colorectal cancer (54.8% male, mean age 68.1)	467 Participants with available blood and stool samples	183 advanced adenoma 67 colorectal cancer 217 without neoplasm

Study	Model of interest	TRIPOD Classification	Study Design	Country	Setting	Participant Description	Outcome	Sample Size for Model	Events
Kim, 2014	1d	2b	Case control development and validation study. Subjects divided into two independent sets.	Korea	Development dataset – patients in which western blot analysis was performed on samples from 81 patients with CRC (and 51 controls) Validation dataset – independent patient cohort of 94 cases and 100 control subjects. Final model uses a combination of both datasets for model development.	Cases and controls Development – 81 cases (mean age 63.16 (SD 10.42)), 51 controls (mean age 50.24 (SD 10.12)) Validation – 94 cases, 100 controls (mean age 62.96 (SD 11.97)) Combined – 132 cases, 151 controls (mean age 49.43 (SD 10.78)). Breakdown by sex not given.	Colorectal cancer	Development – 81 cases, 51 controls Validation – 94 cases, 100 controls Combined – 132 cases, 151 controls	Development – 81 Validation – 94 Combined – 132
Karl, 2008	1d	2b	Case control study/enriched study with stool samples obtained from 2 European multicenter studies.	Europe (Germany)	Study 1 – Preventive screening study at gastroenterology units Study II – Cancer collection study at surgery units	Clinical samples from both multicentre studies were compiled: - Group A comprised the control cohort with 252 patients from study I. - Group B comprised the advanced adenoma cohort containing 113 patients from study I and study II -Group C comprised the CRC cohort with 186 CRC patients from study I and study II. Cancer patients were divided into collective I (no FOBT testing or visible blood in stool), and collective II (no restrictions applied).	Colorectal cancer or Advanced adenoma	Development used CRC collective I and controls. BLR then applied to all samples in collective I (n=101) (controls n=252) to learn a final diagnostic rule and again its thresholds were determined.	252 controls 113 advanced adenomas 186 colorectal cancers

Table 4: Study Characteristics for the eight included studies.

Tripod classification includes; 1a - Development only, 1b - Development and validation using resampling, 2a - Random split-sample development, 2b – Nonrandom split-sample development and validation, 3 – Development and validation using separate data, 4 – Validation study. Model of interest classification for this review includes; 1a - Model development (ideally with internal validation), 1b – Model development with external validation, 1c – External Validation, 1d – Developing/validating a model and then applying as a test, 1e – Applying the risk prediction model as a test (impact study incl. diagnostic accuracy).

3.4 Predictors

Studies assessed a range of predictors including demographic factors which would have been routinely available, novel biomarkers, lab test results and additional lifestyle information from questionnaires. There was some overlap with predictors included in the models; the most frequently included predictors were age and sex. These factors are most likely to be readily available and are well known risk factors for colorectal cancer. Three studies assessed purely demographic factors.[70 71 74] **Table 5** details the included predictors for each of the models developed.

Yen *et al.*[9] and Stegeman *et al.*[16] utilized questionnaires in order to gain additional data on lifestyle factors and other conditions. This is a richer source of data than the routinely available predictors but requires additional data collection. For example, further data on smoking habits and alcohol consumption were obtained along with family history of CRC. Yen *et al.*[9] also carried out additional lab tests; fasting glucose and lipid profile (triglyceride/total cholesterol) in order to obtain further predictors.

Kim *et al.*[72] investigated combining FIT with Calgranulin B (CALB) and age in order to assess the incremental benefit of CALB to a prediction model including the FIT. Tao *et al.*[73] and Karl *et al.*[75] incorporate novel biomarkers into the prediction models. The studies which integrated lab/biomarkers tended to have more of a diagnostic focus.

In terms of how continuous variables were modelled, Stegeman *et al.*[16] treated quantitative variables as continuous in the multivariable model. Restricted cubic splines were used to evaluate the linearity of predictors which led to the square root of the FIT result being used in the multivariable analysis. Lab and biomarkers were in general modeled continuously within the relevant studies.[72 73 75] For Kim *et al.*[72] CALB is modelled as a rank transformed value to accommodate the non-normality, the conversion table is supplied within the supplementary material of the paper.

Categorisation of predictors was commonly used. For example, for the models produced by Yen *et al.*[9] FIT is treated as an ordinal variable using 10 categories and BMI and triglyceride levels are also categorized in the final models. Wieten *et al.*[74] models both age and fecal haemoglobin concentration as 10 unit increases. Finally, Auge *et al.*[71] uses quartiles of FIT

combined with age and sex (log of FIT is used for graphical representation). Categorisation of continuous predictors is not recommended due to the loss of information.[61 77 78]

The events per predictor (EPP) rule of thumb requiring 10 events per predictor[61 79] was met by all studies as far as could be identified in terms of predictors reported for multivariable inclusion. When assessing the EPP for all considered predictors, Stegeman *et al.*[16] mention 13 predictors in the methods, results and **Table 1** of the paper. Based on the 101 events and 13 predictors in this study, this requirement is not quite met with an EPP of 7.77.

3.5 Statistical Analysis

Most of the studies used conventional statistical methods to develop the risk prediction model. Logistic regression was used in five out of the eight studies. Two further studies used a modified version of logistic regression; for example Karl *et al.*[75] used a Bayesian approach to logistic regression and Omata *et al.*[70] used the three-category status of CRN as the dependent variable in a nonlinear ordinal logistic regression. The final study by Yen *et al.*[9] used survival analysis methods (accelerated failure time models) due to staggered invitations to screening.

Predictor selection was not always reported sufficiently in the included studies. Those studies which used univariable screening of variables included Auge *et al*.[71] and Wieten et al.[74] where variables were included if they had a p value of less than 0.05. For Wieten *et al.*[74] sex was also included based on clinician's rationale which reduces some of the data driven selection. Studies which use univariable screening can lead to bias in the predictor effects because although not significant in isolation, a variable may become significant when combined with other factors in a multivariable model. Stegeman *et al.*[14] conducted a previous study which assessed the importance of various predictors for colorectal cancer and included these predictors in the analysis. Studies which had a test accuracy focus as well as a model development focus tended to use all available predictors or test the incremental benefit of including an additional parameter. The model was seen as a method of combining several biomarkers or an additional lab result for an overall result.

Internal validation methods were used by half of the included studies (4/8). Cross validation was used by Kim *et al.*[72] (leave one out cross validation) and Karl *et al*.[75] (100

runs in a Monte-Carlo cross-validation) when developing the Bayesian logistic regression to estimate results for diagnostic performance. Omata *et al.*[70] assesses the internal validity of the model using bootstrapping which showed an optimism of 0.06 but it is not described whether this was for model performance measures or for model parameters. Finally, Stegeman *et al.*[16] used penalized shrinkage to estimate coefficients and to correct the logistic regression model for optimism. Correcting for optimism reduces the overfitting of the model so it is more generalizable to new populations. Other studies do not report any adjustment for optimism explicitly in the text.

Model building strategies were not always described in enough detail to allow replication of the methods. Backwards elimination was used to build a logistic regression model by Stegeman *et al.*[16] using a p value of 0.02 as the removal level. Other studies just included predictors which were significant at a certain p-value (0.05).[9 70 71 73 74]

Two studies used a split sample method to develop their model as well as using cross validation and applying the model in an 'independent dataset'. The final model by Kim *et al.*[72] presented as a model equation is based on a model produced from combining both development and validation datasets. Despite validating the model internally and externally, this final model is based on combining both the validation and development data samples, rather than using the coefficients from the original developed model either adjusted for optimism or recalibrated in the new external dataset. This new model which has been developed needs subsequent external validation. A similar approach is taken for Karl *et al.*[75] where two models are effectively developed. The first uses 2/3 of the CRC Collective I and controls data. Bayes Logistic Regression (BLR) was then applied to all samples from Collective I to develop a final diagnostic rule and thresholds determined. This final rule was then used to predict test results for the CRC Collective II (authors say in an independent patient cohort) and adenomas. The final model developed would need to be both internally and externally validated depending on which version is used.

Only one study had a pure external validation dataset which was used to assess performance of the model (AUC – discrimination) in a new patient population.[9] For model development, patients were those invited to population based screening for colorectal neoplasia using the FIT in Keelung. The external validation set on the other hand used two community based screening programs which included the FIT, in Changhua and Tainan in

Taiwan (geographical validation). All the colorectal neoplasia cases and a random sample of one-tenth of disease-free subjects were selected for external validation. Although this used an external validation dataset, the population is selected in a case control approach with a random sample of 1/10 of disease free subjects and all the colorectal neoplasia cases which may inflate model performance/test accuracy parameters. A minimum of 100 events are recommended in external validation studies.[80]

Only one study used multiple imputation to account for missing variable data,[16] the other studies either did not report any missing data or reduced the sample size sufficiently to allow for complete results (complete case analysis). Reducing the sample size in this way can lead to unreliable associations between predictors and outcomes.

3.6 Model Performance (Discrimination, calibration)

Six of the eight studies reported discrimination measures (AUC ROC) and three out of eight reported calibration. Calibration was most often reported as the Hosmer-Lemeshow statistic, with one study providing a calibration plot for assessment as well as a histogram showing the distribution of risk.[16] Discrimination ranged from 0.676-0.960 for risk adjusted FIT (reported in 6/8 studies) and 0.683-0.902 for FIT only (reported in 4/8 studies). Lab results and biomarkers tended to give higher discrimination values but a significant improvement was also seen when using simple routinely available predictors such as age and sex which are more readily available. Calibration using Hosmer-Lemeshow ranged from 0.276-0.940 for risk adjusted FIT (reported in 3/8 studies). The highest calibration was reported by Stegeman *et al*. with the combination of lifestyle and routine data.

Omata *et al.*[70] just reports model fit in terms of whether the likelihood ratio test showed an improvement over the model including only FIT ($p<0.001$). FIT only is compared to risk adjusted FIT when reporting discrimination in 4 studies.[9 16 72 73] This is also formally tested by looking at the difference in the AUC ROC and reporting the p value for Yen *et al*.[9] (not significant $p=0.62$), Kim et al.[72] (just significant $p=0.049$) and Stegeman *et al*. (significant $p=0.02$). For those studies using FIT or FIT in a model as a comparator, likelihood ratio testing is used to determine if the model has a significantly better fit for Stegeman *et al*.[16] ($p<0.001$), Omata *et al.*[70] ($p<0.001$) and Tao *et al.*[73] ($p=0.002$ for adding one marker, $p=0.0007$ for adding 3 markers).

Different levels of model discrimination are achieved depending on whether lab based predictors are used, routine demographic data or richer questionnaire data. Stegeman *et al.*[16] used a questionnaire to collect additional predictor information and reported an increase in the AUC from 0.69 (for FIT only adjusted for age) to 0.76 (significant). Auge *et al.*[71] included just age and sex along with the FIT result and had an AUC of 0.676 (95% CI: 0.657-0.695) with no comparator to FIT-only. Yen *et al.*[9] used questionnaire plus lab parameters (AUC of 0.83) and compared this to FIT only (AUC of 0.83) and a model containing conventional risk factors only (AUC of 0.66). Tao *et al.* reported an increase from FIT only (AUC 0.683) to FIT plus biomarkers CRP, sCD26 and TIMP-1 (AUC 0.729). Finally, Kim *et al.*[72] compared a FIT only model (adjusted for age) with an AUC of 89.52 with a model with FIT, age and CALB with an AUC of 92.05 (significant difference in AUC).

Model performance results are summarized in **Table 5**.

Study	Predictors considered	Predictors included	How continuous predictors modelled	EPP (events per predictor)	Modelling Method	Predictor Selection	Extra Predictor data	Internal/external validation method and optimism adjustment	Discrimination	Calibration	Other performance measures	Risk Assessment	Risk Presentation
Stegeman, 2014	Predictors mentioned in the methods, results and Table 1: FIT result, age, calcium intake, CRC family history, smoking, BMI, menopausal status, fibre intake, aspirin/NSAID use, red meat intake, sex, alcohol intake, physical activity	FIT result, age, calcium intake, family history, past or current smoking	Restricted cubic splines, square root of FIT result. Quantitative variables treated as such in the multivariable model	For all considered predictors: 101/13 = 7.77 For the final model: 101/5 = 20.2	Logistic Regression	Backwards Elimination 0.20 as removal criterion	FIT plus risk questionnaire	Corrected for optimism by penalized shrinkage – coefficients are estimated with penalized maximum likelihood the optimal penalty factor is determined with the akaike information criterion	AUC FIT only model – 0.69 AUC Risk model – 0.76 Discrimination improved significantly with the risk based model (p=0.02)	Hosmer-Lemeshow test p=0.94 Calibration plot of the risk model	FIT only model (adjusted for age) compared to risk based FIT p<0.001 goodness-of-fit likelihood ratio test NRI 0.054 p=0.073 5 more cases advanced adenoma detected by referring those at highest risk Distribution of risk.	Risk as a probability for advanced neoplasia with a threshold of 0.19	Odds ratios presented and the risk equation is used to generate probabilities.
Yen, 2014 (Results for the model predicting colorectal neoplasia)	Structured questionnaire; dietary habits (alcohol, smoking), family history, and personal history of cancer, type 2 diabetes, hypertension, cerebrovascular disease, cardiovascular disease (hypertension), BMI. Additional lab tests; fasting glucose, lipid profile (triglyceride/total cholesterol), FIT result. Demographic factors; gender	Gender, FIT result, family history of CRC, diabetes mellitus, hypertension, alcohol drinking, smoking, BMI, triglyceride FIT and Gender also assessed.	FIT is treated as an ordinal variable with 10 categories. BMI and triglyceride also categorized. FIT as a continuous variable is assessed but model not reported.	**Development**: 1147/22 = 52.14 (for the final model) **Validation**: 2911/22 = 132.32 (for the final model)	Accelerated Failure Time model	Predictor selection not reported – includes those with a 5% significance level	FIT plus fasting glucose, lipid profile and a structured questionnaire on lifestyle	**External validation:** AUC for the model containing only FIT was 85.6% (95% CI – 84.8-86.4%) Conventional risk factors only 63.6% (95% CI – 62.5-64.7%) FIT plus conventional risk factors 86.1% (95% CI – 85.2-86.9%)	**Development**: AUC for the model containing only FIT was 83.0% (95% CI – 81.5-84.4%) Conventional risk factors only 65.8% (95% CI – 64.2-67.4%) FIT plus conventional risk factors 83.5% (95% CI – 82.1-84.9%)	-	Adding conventional risk factors to the model with FIT only did not make a significant contribution (p=0.62) (AUC)	The probability of having incident colorectal neoplasia.	Risk equations presented. The regression coefficients from the AFT models were used as the clinical weights in external validation.
Omata, 2011	FIT result, age, BMI and sex	BMI, age, sex, FIT	Nonlinear effect was assessed by testing linear effect for all variables – method unclear	For predictors in the final model Events per predictor for SN 69 AAP + 8 CRC 77/4 = 19.25	Nonlinear Ordinal Logistic Regression using three-category status of CRN as dependent variable	Not reported. In unadjusted analysis for SN there was significant difference in age, sex, BMI and QTFIT between disease categories	CRN data obtained from general health checkup	Internal validity of prediction model investigated by bootstrapping. Optimism 0.06. Further details not given.	-	-	Presents 10 profiles with probability of adenomatous polyp and significant neoplasia and significant neoplasia Likelihood Ratio	Prediction of SN	Nomogram for predicting AP or SN and SN

				SN or AP SN = 77 AP = 393 (including 69 cases of AAP) 470/4 = 117.5							test showed an improvement over the model including only QTFIT (p<0.001)		
Auge, 2014	Sex, age, FIT quartiles, referral hospital, city district and primary care centre	FIT result, age, sex	Log FIT for graphic representation but quartiles of FIT used for multivariable model. Age is categorized into 2 groups: 50-59 60-69	For all considered (6 variables) 1441/6 = 240.17 Final model 1441/3 = 480.33	Logistic regression	Included variables with a p value of less than 0.05 in the multivariable model	FIT plus demographic characteristics	No internal validation/adjustment for optimism	AUC ROC 0.676 (95% CI: 0.657-0.695)	Hosmer-Lemeshow test p=0.312	-	Logistic regression model was used to delineate 16 risk categories for advanced colorectal neoplasia	16 risk categories associated with different probabilities for advanced colorectal neoplasia These categories were then classed into 3 risk levels (arbitrarily) with different probabilities according to the corresponding PPV. Up to 30% - low risk 31%-50% - average risk >50% - high risk
Wieten, 2016	Sex, age, socioeconomic status, FIT concentration	Age, gender, FIT	Age (per 10 year increase) Fecal Hb Concentration (per 10 µg Hb/g increase)	For all considered variables (parameters) 193/5 = 38.6 For final model 193/3 = 64.33	Logistic regression	Univariable screening (p<0.05) plus clinicians rationale	FIT plus demographic characteristics	No internal validation/adjustment for optimism	-	Hosmer-Lemeshow goodness of fit: p=0.276	-	Predicted probability of having AN per screenee who had a positive FIT and subsequent colonoscopy	Predicted probabilities of having AN per screenee were depicted in a figure with age as a continuous variable
Tao, 2012	C-reactive protein (CRP), serum CD26 (sCD26), complement C3a anaphylatoxin, tissue inhibitor of metalloproteinases I (TIMP-1) – measured by ELISA tests and FIT and guaiac FOBT also	FIT, CRP, sCD26 and TIMP-1	Assumption modeled continuously blood markers as (ng ml^{-1}) and FIT concentration as µg Hb/g.	For all considered predictors (advanced adenoma results) 183/5 = 36.6	Logistic regression	Included predictors whose levels showed statistically significant differences in CRC cases versus	FIT plus blood based inflammatory markers (lab results)	No internal validation/adjustment for optimism	FIT plus TIMP-1 AUC = 0.710 FIT plus CRP, sCD26 and TIMP-1 AUC = 0.729 FIT only	-	(Advanced adenoma) Model fit was significantly improved adding TIMP-1 to FIT result (p=0.002) or all three markers (FIT,	Predicted probability of advanced adenoma	Model not presented

	performed.			Final model 185/4 = 46.25		participants free of neoplasms			AUC = 0.683		CRP, sCD26 and TIMP-1) (p=0.0007) using likelihood ratio tests		
Kim, 2014	Fecal CALB (calgranulin B), FOBT, age	Fecal CALB, FOBT, age	CALB is a rank transformed value (conversion table supplied in supplementary material). Their rank was used in to accommodate the non-normality. Assume FOBT and age are continuously modeled from risk equation	Development: 81/3 = 27 Validation: 94/3 = 31.33 Combined development: 132/3 = 44	Logistic regression	The aim of the paper was to assess the incremental benefit of CALB to a prediction model including FOBT. The model was also adjusted for age due to imbalances between CRC patients and controls in both the sets	FIT plus CALB (lab result)	Internal validation - LOOCV (leave-one-out cross validation) External validation using an independent set (non-random split sample development and validation) **Final model presented combines both development set and validation set.** *Model (age + FIT)* AUC – 0.9017 Partial AUC – 0.0699 *Model (age + FIT + CALB)* AUC – 0.9282 Partial AUC – 0.0748	**Development dataset**: *Model (age + FIT)* AUC – 89.52 Partial AUC – 6.65 *Model (age + FIT + CALB)* AUC – 92.05 Partial AUC - 7.02 **Development LOOCV dataset**: *Model (age + FIT)* AUC – 87.78 Partial AUC – 5.62 *Model (age + FIT + CALB)* AUC – 89.81 Partial AUC – 5.70 **Validation dataset**: *Model (age + FIT)* AUC – 90.65 Partial AUC – 7.34 *Model (age + FIT + CALB)* AUC – 92.74 Partial AUC – 7.71	-	For the development dataset. Reclassification improvement p-value of RI in CRC patients and controls from the model using FOBT adjusted for age 0.0013 and for the full model 0.0173. P-value of NRI was 0.0001. Difference in AUC ROC for incremental benefit of CALB p-value 0.0499	Predicted probability of CRC. Cut point is at a specificity closest to 90%	Risk equation for the model combining both datasets is presented
Karl, 2008	Hemoglobin, Hemoglobin-haptoglobin, S100A12, TIMP-1, Calprotectin, CEA	Model Combination 1: Hemoglobin-haptoglobin, S100A12 Model Combination 2: Hemoglobin-haptoglobin,	Continuous	Collective 1 (n=101 CRC) Controls (n=252) 2/3s used for development. 0.66*101 = 67.33 68/6 = 11.33	Bayes Logistic Regression	Six marker candidates were evaluated alone or in combination for the detection of CRC in stool samples using BLR. BLR 'selected' the combinations.	FIT plus biomarkers	Results of BLR were evaluated by 100 runs in a Monte-Carlo cross-validation design applied on CRC collective I and controls. Using CRC collective II they validated the results in an independent patient cohort.	AUC 0.95 for S100A12 alone to 0.96 for marker combinations (Haemoglobin-Haptoglobin plis S100A12), (Haemoglobin-Haptoglobin plus S100A12 plus TIMP-1)	-	-	A threshold on the estimated posterior case probabilities was determined on the controls of the training set to achieve an	Model not presented

		S100A12, TIMP-1		(all considered predictors)								apparent specificity of 95% or 98% for the multivariable diagnostic rule. BLR then was applied to all samples in collective I to learn a final diagnostic rule and again its thresholds were determined.	

Table 5: Model Characteristics for the eight included studies. BLR = Bayesian logistic regression.

3.7 Other Performance Measures (Net Reclassification Improvement)

Two studies reported classification related parameters.[16 72] For Net Reclassification Improvement (NRI), reclassification tables are produced separately for both participants with events, and those without events and then the correct movement in categories is quantified. For example, an improvement in reclassification occurs when participants with the outcome are re-classified into higher risk groups (move upwards) or when a participant without the outcome moves into lower risk groups (downwards).[81] This reclassification metric allows insight into the added value of predictors to a model.[82]

The NRI for the FIT only compared to risk based FIT was 0.054 with a p value of 0.073.[16] Five more cases of advanced adenoma were detected by referring those at highest risk. For Kim *et al.*[72] the p value for the NRI was 0.0001 when comparing the incremental benefit of calgranulin b to a model with just FIT adjusted for age, indicating a significant improvement in reclassification.

3.8 Individualised Risk Prediction: Presentation and Application

Only two studies published the overall risk equation for use in other populations.[9 72] The model presented by Kim *et al.*[72] was based on combining both the development and validation sets and would therefore need internal/external validation to assess its performance in another patient population. Yen *et al.*[9] provided model equations for all nine models produced, including the parameterisation of the baseline hazard. The same weights from these models are then applied in the external validation assessment. One study presents the risk model as a nomogram, which needs external validation.[70] When a nomogram is presented it is likely to not use the full version of regression coefficients for probability assessment. Three studies present the odds ratios for the multivariable model but the intercept would be required for the full prediction model if converting odds ratios to weight parameters.[16 71 74]

Auge *et al.*[71] uses the model with FIT result, age and sex to produce probabilities for advanced colorectal neoplasia. The probabilities were used to delineate 16 risk categories and 3 risk levels arbitrarily based on corresponding PPVs. Up to 30% PPV is assigned as low

risk, 31-50% is assigned as average risk and over 50% is high risk. The authors suggest these categories could be used to prioritise individuals for colonoscopy.

3.9 Test Accuracy

Four of the studies included an assessment of the risk prediction model applied as a test and report the sensitivity at a set specificity or recall rate using the same level achieved by the FIT/FOBT.[16 72 73 75] Sensitivity ranged from 21.9% to 88% for risk adjusted FIT at a range of set specificities from 90-97.7%. FIT only sensitivity ranged from 19.7% to 82% at specificities from 90-97.7%. The highest sensitivities were seen in studies combining further lab test results including calgranulin B, S100A12 and TIMP-1. The relative improvement in sensitivity from FIT only to risk-adjusted was similar for a study combining questionnaire data (32% to 40%, an 8% increase). Outcomes for these models were advanced adenoma,[73] advanced colorectal neoplasia,[16] and colorectal cancer[72 75].

The sensitivity of risk-adjusted FIT (the model including FIT with other predictors) along with FIT only is plotted in ROC space for these 4 studies in **Figure 2**. Risk-adjusted FIT has a higher sensitivity at similar specificity for all these studies. Stegeman *et al.*[16] set the recall rate the same (same specificity as FIT only) giving a risk threshold of 0.19; this is then used to define TPs, TNs, FPs, FNs. This is the only study which reports the 2 by 2 data required for assessment of test accuracy (See Forest Plot, **Figure 3).**

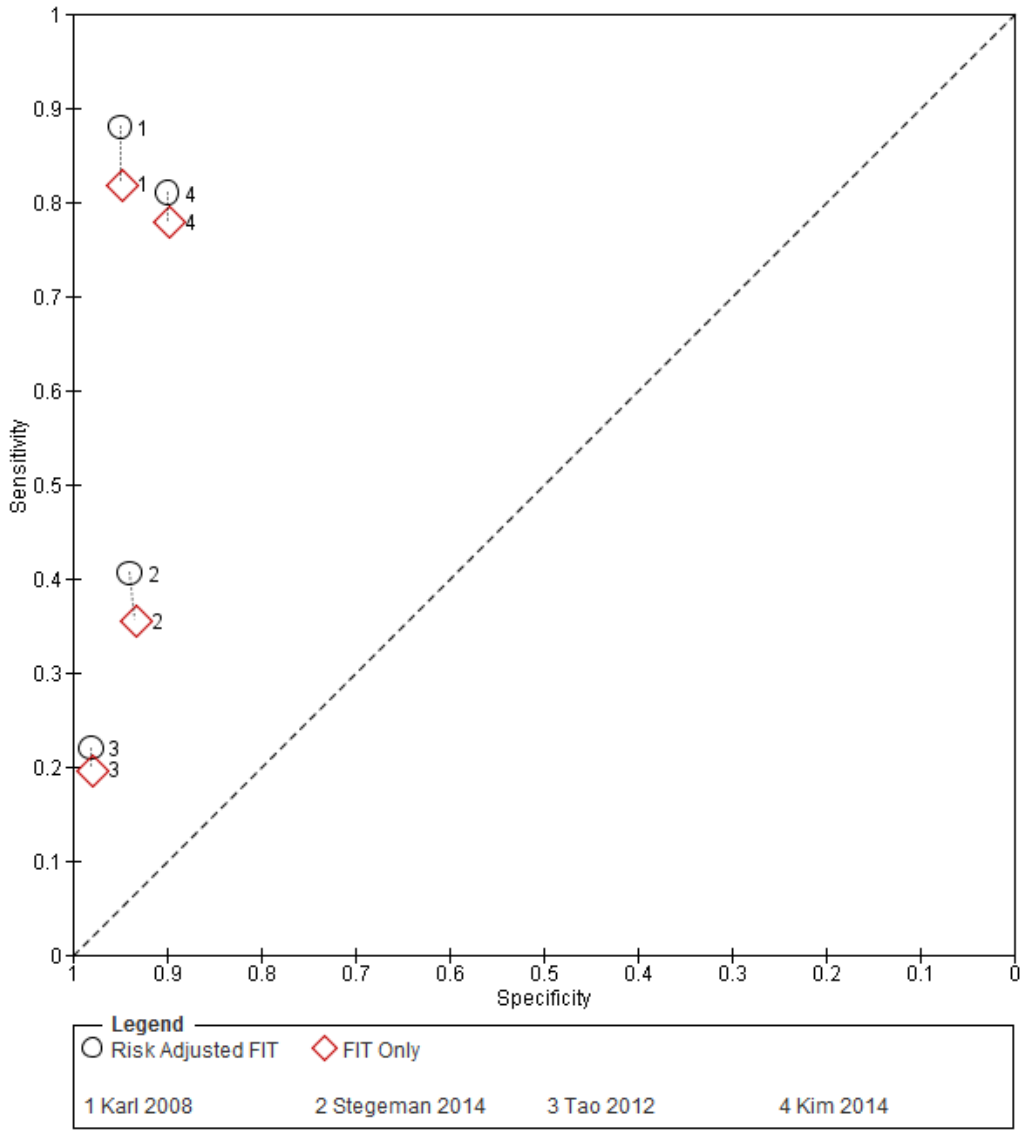


Figure 2: Sensitivity and Specificity plotted in ROC space for the studies which applied the risk prediction model as a test. These studies compare the test accuracy of the model to FIT only. Outcomes differ between the studies (Stegeman 2014 advanced neoplasia), (Karl 2008 colorectal cancer results using the cross-validated model at 95% specificity), (Kim 2014 colorectal cancer results using the model combining both development and validation sets), (Tao 2012 advanced adenoma).

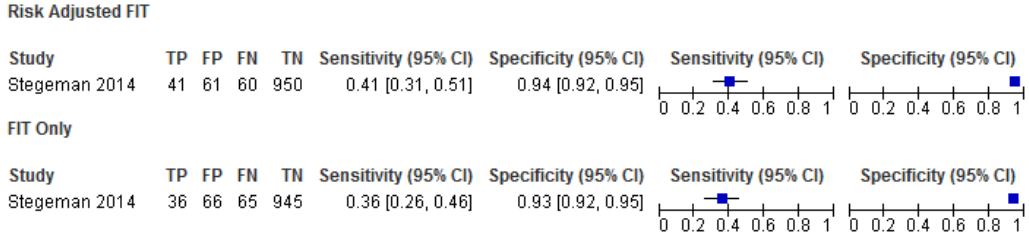


Figure 3: Forest plot of sensitivity and specificity for Stegeman et al. Other studies did not report or it was not possible to derive 2 by 2 data.

Author Year	Type and Brand of FIT	Threshold	Reference Standard	2 by 2	Sensitivity (95% CI)	Specificity (95% CI)	AUC ROC	PPV/NPV
Stegeman, 2014	OC Sensor	50 ng/ml 10 µg Hb/g	Colonoscopy	FIT plus risk - TP (41), TN (950), FN (60), FP (61) FIT only - TP (36), TN (945), FN (65), FP (66)	40% - risk based model, 32% for FIT only	93% specificity for FIT only and FIT plus risk.	AUC of FIT only model - 0.69 AUC of Risk based model - 0.76 (AUCs significantly different p=0.02)	FIT plus risk - 40.2% (PPV); 94.1% (NPV). FIT only - 35.3% (PPV); 93.6% (NPV) (calculated)
Yen, 2014 **(Results for the model predicting colorectal neoplasia)**	OC-Sensor (Eiken)	100ng/ml 20 µg Hb/g	Colonoscopy following referral for screen detected cancers or linkage to the Cancer Registry	-	-	-	**Development:** 83.5% (95% CI: 82.1%, 84.9%) **External Validation:** 86.1% (95% CI: 85.2%, 86.9%)	-
Omata, 2011	OC-Micro Instrument (Eiken Chemical Co., Tokyo, Japan)	For AP: 25ng/ml For SN: 25ng/ml For CRC: 50ng/ml	Colonoscopy and pathological findings of biopsy specimens	No results for the risk prediction model applied as a test.	-	-	-	-
Auge, 2014	OC-SENSOR DIANA	20 µg Hb/g	Colonoscopy	-	-	-	AUC ROC 0.676 (95% CI: 0.657-0.695)	Risk categories created using PPV. Low risk (30%>), average risk (31-50%) and high risk (>50%)
Wieten, 2016	OC-sensor Micro analyser	10 µg Hb/g	Colonoscopy	No results for the risk prediction model applied as a test.	-	-	-	-
Tao, 2012 (results for advanced adenoma)	RIDASCREEN Haemoglobin	Sensitivities were calculated at cutoff points yielding the level of specificity observed for gFOBT (97.7%) (the test applied at the time in Germany) – 24 µg Hb/g	Colonoscopy	-	Sensitivity 21.3% for TIMP plus FIT Sensitivity 21.9% for CRP, sCD26 and TIMP-1) plus FIT Sensitivity FIT alone 19.7% All at cutoff points yielding 97.7% specificity	97.7% Specificity	FIT plus TIMP-1 AUC = 0.710 FIT plus CRP, sCD26 and TIMP-1 AUC = 0.729 FIT only AUC = 0.683	-
Kim, 2014	OC Sensor kit	Analytical cut off positivity = 100ng Hb/ml 20 µg Hb/g Sensitivity at a specificity closest to 90%	CRC diagnosed by colonoscopy and histopathology	-	**Development dataset**: *Model (age + FIT)* -Sensitivity – 75.31% -Specificity – 90.2% *Model (age + FIT + CALB)* -Sensitivity – 83.95% -Specificity – 90.2% **Development LOOCV dataset**: *Model (age + FIT)*	90% Specificity	**Development dataset**: *Model (age + FIT)* AUC – 89.52 Partial AUC – 6.65 *Model (age + FIT + CALB)* AUC – 92.05 Partial AUC - 7.02 **Development LOOCV dataset**: *Model (age + FIT)*	-

					-Sensitivity – 75.31% -Specificity – 90.2% *Model (age + FIT + CALB)* -Sensitivity – 82.72% -Specificity – 90.2% **Validation dataset**: *Model (age + FIT)* -Sensitivity – 79.79% -Specificity – 90% *Model (age + FIT + CALB)* -Sensitivity – 79.79% -Specificity – 79.79% **Final model presented combines both development set and validation set.** *Model (age + FIT)* -Sensitivity – 77.71% -Specificity – 90.07% *Model (age + FIT + CALB)* -Sensitivity – 80.57% -Specificity – 90.07%		AUC – 87.78 Partial AUC – 5.62 *Model (age + FIT + CALB)* AUC – 89.81 Partial AUC – 5.70 **Validation dataset**: *Model (age + FIT)* AUC – 90.65 Partial AUC – 7.34 *Model (age + FIT + CALB)* AUC – 92.74 Partial AUC – 7.71 **Final model presented combines both development set and validation set.** *Model (age + FIT)* AUC – 0.9017 Partial AUC – 0.0699 *Model (age + FIT + CALB)* AUC – 0.9282 Partial AUC – 0.0748	
Karl, 2008	Hemoglobin-haptoglobin (RIDASCREEN Hemoglobin-Haptoglobin)	A threshold on the estimated posterior case probabilities was determined on the controls of the training set to achieve an apparent specificity of 95% or 98% for the multivariable diagnostic rule. BLR then was applied to all samples in collective I to learn a final diagnostic rule and again its thresholds were determined.	Study 1: Colonoscopy Study 2: Diagnosis of CRC confirmed by pathologic staging of each patient	-	**Sensitivity at a specificity of 95%** Median sensitivities from cross validation CRC collective I. -Hemoglobin-Haptoglobin: 82% -S100A12 + hemoglobin-haptoglobin: 88% -S100A12 + hemoglobin-haptoglobin + TIMP-1: 88% **Sensitivity at a specificity of 98%** -Hemoglobin-Haptoglobin: 73% -S100A12 + hemoglobin-haptoglobin: 79% -S100A12 + hemoglobin-haptoglobin + TIMP-1: 82% Median sensitivities from CRC collective II -Hemoglobin-Haptoglobin: 85% -S100A12 + hemoglobin-haptoglobin: 88% -S100A12 + hemoglobin-haptoglobin + TIMP-1: 88% **Sensitivity at a specificity of 98%** -Hemoglobin-Haptoglobin: 78% -S100A12 + hemoglobin-haptoglobin: 82% -S100A12 + hemoglobin-haptoglobin + TIMP-1: 86%	Sensitivity at a preset specificity of 95% and 98% investigated.	0.95 for S100A12 alone to 0.96 for marker combinations (Haemoglobin-Haptoglobin plus S100A12), (Haemoglobin-Haptoglobin plus S100A12 plus TIMP-1)	-

Table 6: Test Characteristics for the eight included studies.

3.10 Risk of Bias

3.10.1 PROBAST

The risk of bias of the models in the included studies was assessed using an early version of PROBAST for both model development and model validation (**Table 7)**. In the context of risk prediction models, risk of bias assesses the extent of unbiased estimates of model performance for intended use and target population. This tool considers the following domains: Study participants, predictors, outcome, sample size and missing data and statistical analysis. The applicability assessments are related to the first three of these domains.

All studies had an 'overall judgement' rated as a 'high risk' of bias (**Figure 4**). The two domains with the greatest concerns included 'statistical analysis' where 90% of models (9/10) had a high risk of bias and 'sample size and missing data' where 60% of models (6/10) had a high risk of bias and the remaining 40% (4/10) had an unclear risk of bias. For statistical analysis of prediction models, the methods were not often adequately reported, there was a lack of internal validation among model development studies and there was rarely an adjustment for optimism. Further to this, only one study could be considered as using external validation to assess the model,[9] and in this case it was a case-control validation sample which may bias model performance results. Stegeman *et al*.[16] was considered the most methodologically sound study in terms of statistical analysis which was rated as a low risk of bias (**Table 7**). The model was developed using penalised shrinkage methods and model development was fully reported.

Often due to the nature of screening studies, sample size and missing data were either at high risk of bias or not enough information was provided to make an assessment. Sample populations in screening studies also tend to have limited populations/reduced sample size based on screening test and reference standard uptake. Further to this, the included studies have the added constraint of additional predictor information to collect which may not always be completed by all individuals. This would be less of a problem for routine demographic data compared to additional lab test results. Stegeman *et al.*[16] was the only study which used multiple imputation for missing predictor data, other studies just used a complete case analysis.

Half of the studies had a high concern regarding applicability (extent to which the model matches the review question). This was mostly down to the case-control design or enriched sample population.[72 73 75] Karl *et al.*[75] and Tao *et al.*[73] used data from two different sample populations. The timing of predictors and outcomes differed for the two sample populations as well as the level of underlying risk. Omata *et al.*[70] used a sample population which was 70% male and some individuals had a history of CRN putting them at higher risk.

Study	RISK OF BIAS						APPLICABILITY CONCERNS				USABILITY
	STUDY PARTICIPANTS	PREDICTORS	OUTCOME	SAMPLE SIZE & MISSING DATA	STATISTICAL ANALYSIS	OVERALL JUDGEMENT OF BIAS	STUDY PARTICIPANTS	PREDICTORS	OUTCOME	OVERALL JUDGEMENT OF APPLICABILITY	USABILITY OF THE MODEL
Stegeman, 2014	+	+	+	−	+	−	+	+	+	+	N
Yen, 2014 (Development)	?	?	−	?	−	−	+	+	+	+	N
Yen, 2014 (Validation)	−	?	?	?	−	−	?	+	?	?	N
Omata, 2011	−	+	+	−	−	−	−	+	+	−	N
Auge, 2014	+	+	+	−	−	−	+	+	+	+	N
Wieten, 2016	+	+	+	−	−	−	+	+	+	+	N
Tao, 2012	−	−	−	−	−	−	−	−	−	−	N
Kim, 2014 (Development)	−	+	+	?	−	−	−	+	+	−	N
Kim, 2014 (Validation)	−	+	+	?	−	−	−	+	+	−	N
Karl, 2008	−	−	+	−	−	−	−	−	−	−	N

+ Low Risk − High Risk ? Unclear Risk

Table 7: Tabular display of PROBAST assessments for the eight included studies.

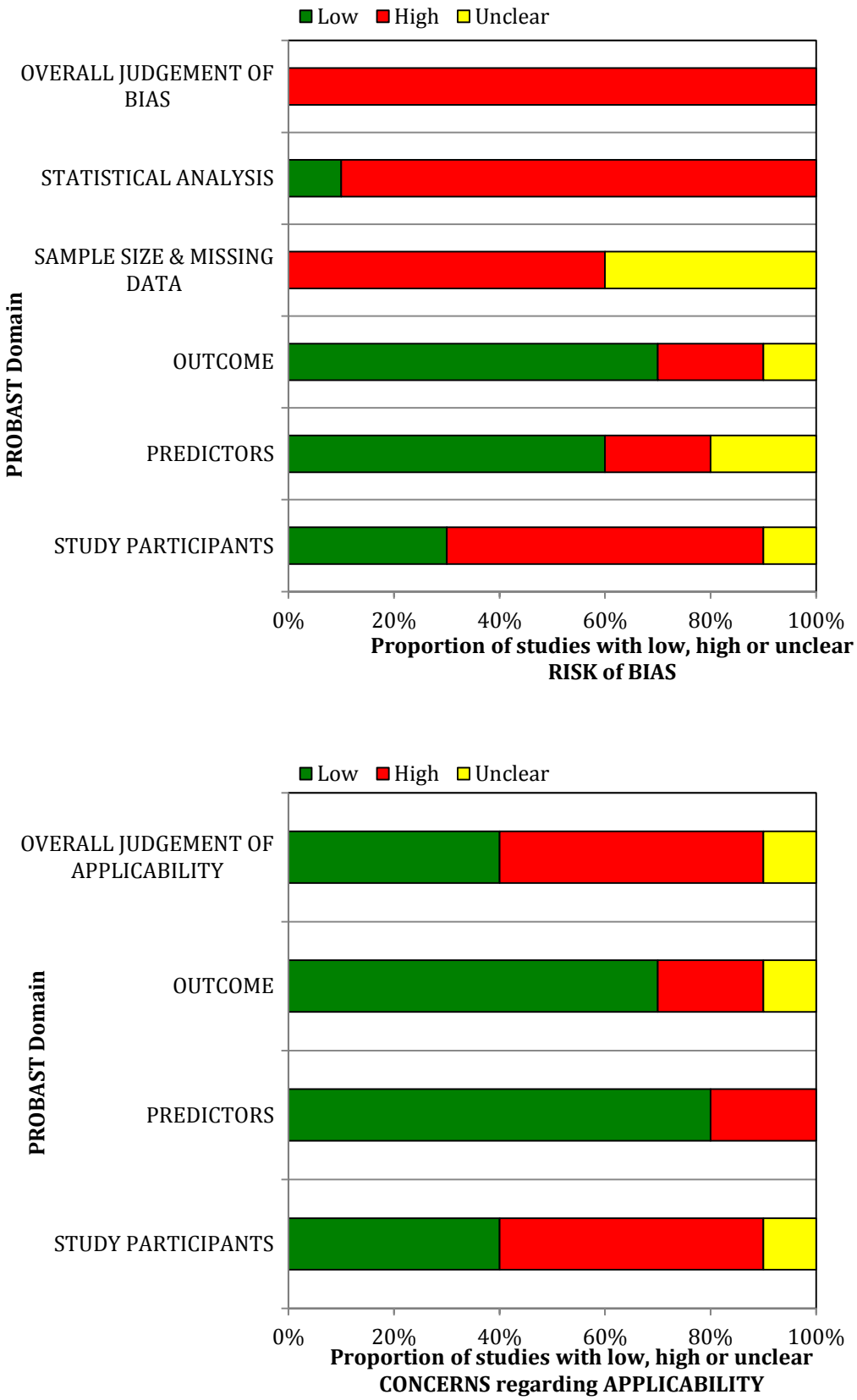


Figure 4: Graphical summary of PROBAST assessments for all included studies. Top – displays the proportion of studies with low, high or unclear Risk of Bias, Bottom – displays the proportion of studies with low, high or unclear concerns regarding applicability

3.10.2 QUADAS-2

Four of the studies also assessed the diagnostic accuracy of the model and reported test accuracy measures of sensitivity, specificity as well as the AUC ROC (interchangeable with model performance). A tailored QUADAS-2 assessment was carried out for this purpose and is summarized in **Table 8**.

'Flow and timing' and 'patient selection' was rated as high risk of bias for 3 out of the 4 studies for similar reasons identified by PROBAST (**Figure 5**). These reasons included the nature of screening studies in relation to uptake of the screening test and reference standard, case control study designs or enriched sample populations. The index test was also rated as high risk of bias for half the studies (when applying the model as the test). This was due to the timing of predictor collection for some participants being after colonoscopy for stool samples.[73 75] With regard to applicability there are again similar issues with patient selection as identified by PROBAST along with the timing of predictor collection for the index tests. The review question is assessing studies where the FIT and predictor information are assessed before colonoscopy.

	RISK OF BIAS				APPLICABILITY CONCERNS		
Study	**PATIENT SELECTION**	**INDEX TEST**	**REFERENCE STANDARD**	**FLOW AND TIMING**	**PATIENT SELECTION**	**INDEX TEST**	**REFERENCE STANDARD**
Stegeman, 2014	☺	☺	☺	☹	☺	☺	☺
Tao, 2012	☹	☹	?	☹	☹	☹	☺
Kim, 2014	☹	☺	☺	?	☹	☹	☺
Karl, 2008	☹	☹	☺	☹	☹	☹	☹

☺Low Risk ☹High Risk ? Unclear Risk

Table 8: Tabular display of QUADAS-2 Assessments for the 4 studies including a test accuracy component

Finally, the role of the sponsor for Karl *et al.*[75] could also be considered at high risk of bias since the study aims to identify new screening markers and combinations to determine if sensitivity can be improved and all authors are employees of Roche Diagnostics GmbH who funded the study. This could bias test accuracy measures by assessing a cutpoint which provides good performance.

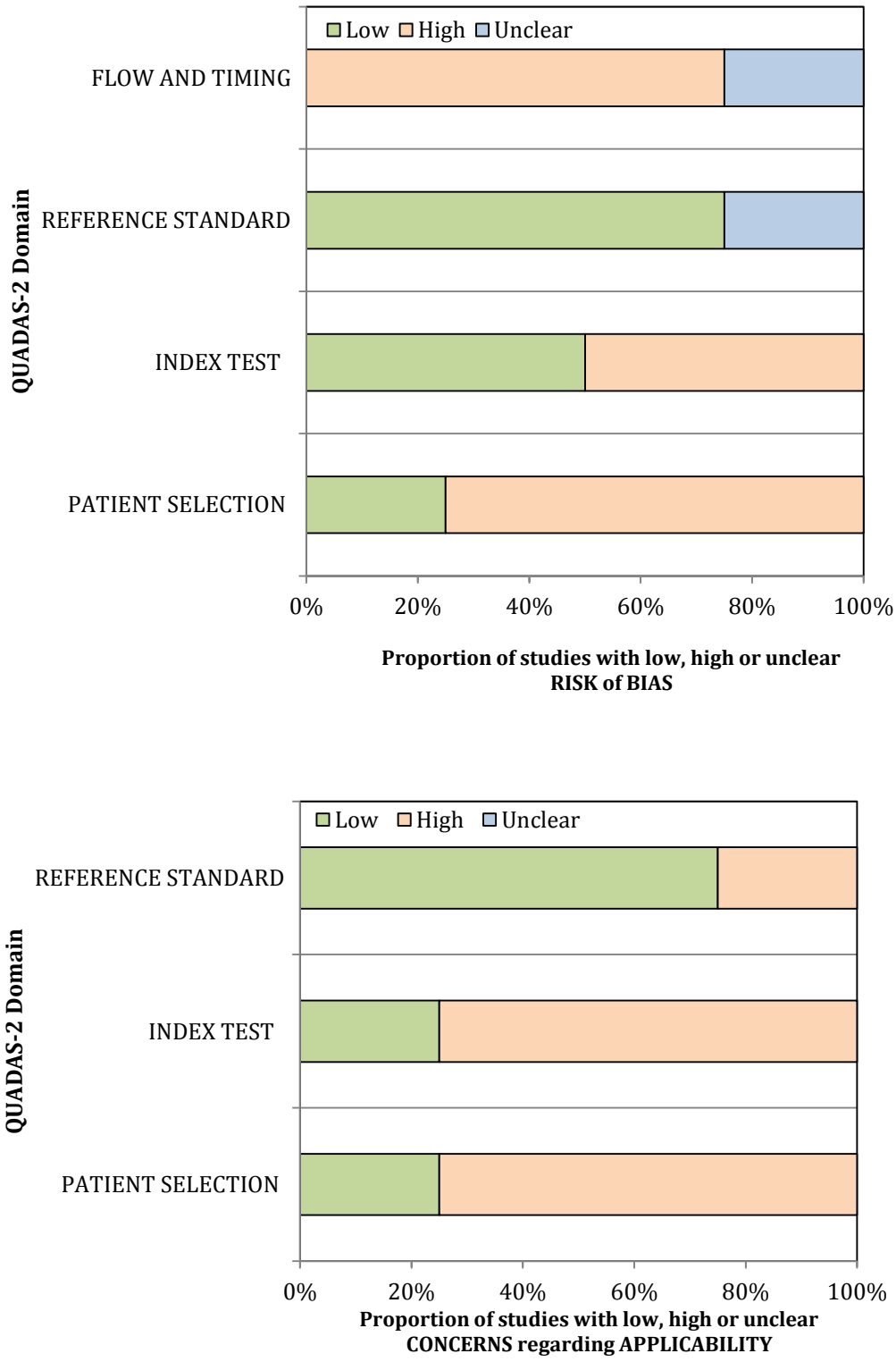


Figure 5: Graphical summary of QUADAS-2 assessments for the 4 studies including a test accuracy component. Top – displays the proportion of studies with low, high or unclear Risk of Bias, Bottom – displays the proportion of studies with low, high or unclear concerns regarding applicability.

4.0 DISCUSSION

4.1 Statement of principal findings

In this systematic review of risk prediction models which combine the FIT result for colorectal cancer screening referral decisions, eight full text studies were identified which develop independent risk prediction models. Predictors used in the models included demographic characteristics, biomarkers and lab results as well as additional information obtained from questionnaires. The most consistently used predictors across the models were sex and age, possibly due to their availability and known association with colorectal cancer.

A meta-analysis of risk prediction model performance and test accuracy was not possible due to the underlying heterogeneity between studies and so a narrative synthesis was performed using summary of results tables. Discrimination assessed as the AUC ROC ranged from 0.676-0.960 for risk adjusted FIT (reported in 6/8 studies) and 0.683-0.902 for FIT only (reported in 4/8 studies). Lab results and biomarkers tended to give higher discrimination values but a significant improvement was also seen when using simple routinely available predictors such as age and sex. High performing models also included lifestyle information such as smoking and alcohol consumption as well as family history of cancer. Lab test results associated with greater model performance included TIMP-1, CEA, sCD26 and calgranulin B. Further evidence is required to confirm which biomarkers and other predictors should be included in a risk based screening model for FIT.

Calibration using the Hosmer-Lemeshow statistic ranged from 0.276-0.940 for risk-adjusted FIT (reported in 3/8 studies) and calibration plots were presented in just one study. The highest calibration was reported by Stegeman *et al.* This may be a reflection of the more rigorous model development procedure used by this study to generate more accurate predictions/results.

Where test accuracy measures were included (4/8 studies) sensitivity ranged from 21.9% to 88% for risk adjusted FIT at a range of set specificities from 90-97.7%. FIT-only sensitivity ranged from 19.7% to 82% at the same specificities stated previously. The highest sensitivities were seen in studies combining further laboratory test results including

calgranulin B, S100A12 and TIMP-1. The relative improvement in sensitivity from FIT only to risk-adjusted was just as much for a study combining questionnaire data (32% to 40%, an 8% increase). Studies focusing more on test accuracy metrics often showed higher sensitivity for FIT only (and risk-adjusted) due to study design and quality. Studies which had a test accuracy focus as well as a model development focus tended to use all available predictors or test the incremental benefit of including an additional parameter. The model was seen as a method of combining several biomarkers or an additional lab result for an overall test result.

Most studies used standard statistical methodologies to build the risk prediction model; 5/8 studies used logistic regression, two further studies used a modified version of logistic regression and 1/8 used accelerated failure time models (survival analysis). Half of the studies considered some form of internal validation. Model building strategies were not always described in enough detail to allow replication of the methods.

Risk prediction models were either presented as the full equation, a Nomogram or by creating risk categories. Two of the 8 studies present the full risk equation, although for one of these studies this would need subsequent internal and external validation. The risk-based categories produced by one of the studies was based on the probabilities from the logistic regression equation and then the subsequent PPV to assign three risk categories. Many of the studies just present the odds ratios only with no intercept or baseline hazard parameterization.

Domains marked consistently at a high risk of bias using PROBAST included ‘statistical analysis’ where 90% of studies had a high risk of bias and ‘sample size and missing data’ where 60% of studies had a high risk of bias. For statistical analysis of prediction models, the methods were not often adequately reported, there was a lack of internal validation among model development studies and there was rarely an adjustment for optimism. Statistical methodology was rated as low risk of bias for just one of the studies.[16] Half of the studies had a high concern regarding applicability. This was usually down to the case-control design or enriched sample populations.

For studies which had a test accuracy component, the QUADAS-2 tool further identified the ‘index test’ domain as being at high risk of bias for three out of the four studies. This was

due to the timing of predictor collection for some participants being after colonoscopy for stool samples.

4.2 Strengths and weaknesses of the study

This systematic review summarises all the available evidence of risk prediction models which combine the FIT result for colorectal cancer screening referral decisions. Searches included grey literature as well as ongoing study searches to capture all available evidence. In addition, the review uses recently developed and published tools for risk prediction models: CHARMS, TRIPOD and a piloted early version of PROBAST.[27 61]

A limitation of the review is that a quantitative synthesis could not be performed due to the underlying heterogeneity. However, the review finds some evidence for improved model performance and test accuracy using a risk-adjusted approach. Further to this, case control/enriched sample studies are included in this review which can inflate test accuracy and bias model performance. Useful information however can be obtained from these studies regarding potential predictors, how the models are presented for application and statistical modelling procedures.

Sample populations in screening studies also tend to have limited populations/reduced sample size based on screening test and reference standard uptake. Included studies have the added constraint of additional predictor information to collect which may not always be completed by all individuals. The applicability of the models therefore needs to be inspected for use in an average risk screening population, if the study for instance only includes responders to screening and those with a positive FIT. The use of these models however could be as a further triage if an individual has a positive FIT to decide on colonoscopy resource use.

Studies reporting test accuracy parameters had more of a diagnostic accuracy focus over model performance and therefore did not focus fully on model development reporting and often did not fully report the methods used to combine FIT with other biomarkers.

4.3 Strengths and weaknesses in relation to other studies

The conclusions relating to reporting and issues in statistical methodology are similar in this review compared to those found in other recent risk prediction reviews.[83 84] Improved reporting of risk prediction studies is likely to correlate with the publication of risk prediction model guidelines such as TRIPOD (published in 2015).[61] A systematic review which synthesized models which predict future risk of colorectal cancer identified a discrimination range of between 0.71 to 0.78 compared to the wider range reported by this review of between 0.676 to 0.960. This is likely due to the differences in outcome definitions, statistical methodologies and since this included apparent performance as well as cross validated and external performance.

This review did not include all possible methods of combining FIT with other risk predictors, the focus was on risk prediction models. The FIT can be combined sequentially, at different time points and in different populations (negative versus positive test results only) or the predicted risk considered for surveillance strategies. A model for predicting advanced colorectal neoplasia in FIT negative people only was recently developed with an intended use of prioritising patients for colonoscopy.[85] For example, some studies considered using a scoring system as a triage (The Asia-Pacific Colorectal Screening score) and then applying the FIT separately.[86 87] The current review however focused on risk prediction models which combined all the information within one model/package. The timing of the risk prediction model would be around the time of the FIT to determine participants who are at highest risk of colorectal cancer for colonoscopy referral.

Furthermore, other studies which use risk prediction models focus on symptomatic populations only and these types of studies were excluded from this review.[88] The NHS Bowel Cancer Screening Programme treats symptomatic and screening participants separately using different computer systems and referral criteria which is why these models were not included in this review. In addition, risk prediction models for CRC in symptomatic patients has been considered in a recently published systematic review.[40]

There are many risk prediction models which consider routine predictors only and do not combine the FIT result.[20 21 89] These were not included as this research investigated the added value of modifying the screening test with further factors in more of a personalised

screening approach. These models also tend to have lower discrimination on their own when not including the FIT.

4.4 Practical Implications

From the studies included in this review there is some evidence to suggest risk adjusted FIT performs better than FIT only. Both demographic characteristics as well as lab test results showed an improved performance in discrimination and test accuracy parameters. Demographic characteristics such as age, sex and other socioeconomic data such as IMD and BMI are likely to be readily available on GP records or screening systems requiring no additional data collection. Lab tests although they improve sensitivity are subject to additional issues such as cost as well as uptake/completion by the target population. If blood tests are taken for general health checkups then this sort of lab data could be used as another source of predictor information. Ideally this would be made available in electronic health records.

Based on PROBAST assessment, none of the models were considered as usable in the target population and screening context. This was mainly due to a lack of external validation and lack of reporting the final model and statistical methodology. Yen *et al.*[9] was the only model which used external validation but this study used a case control participant selection strategy for this purpose.

Application of models could be at the screening programme level and use in-built risk calculators within the screening systems to decide on whether someone is risk positive. Furthermore, risk information can be presented to potentially help informed decision making and increase uptake of the reference standard as well as the screening test.[90 91] If an individual has a lack of perceived risk this can affect uptake.

4.5 Future research

Predictors which were most commonly included in the final models included demographic characteristics age and sex and so these predictors should be considered for inclusion in future model development studies. Routinely available predictors are available from most

screening programme IT systems such as the BCSS in the NHS. These factors could be investigated without requiring additional data collection in a model development study.

With the recent publication of the TRIPOD guidelines, future risk prediction modelling studies should consider these criteria when developing a new model and for reporting. This will help to improve the quality of the study, provide reproducibility of the model development process and allow the model to be validated externally or applied in an impact study.

Experts identified two studies which use machine learning approaches including decision trees with cross validation techniques.[68 69] This model included the guaiac FOBT and various lab parameters and showed that in comparison to gFOBT alone the number of CRC cases increased from 170 to 365 (AUC of 0.82) demonstrating that such an approach could improve the performance of other screening tests. In addition, the model can detect CRC throughout the colon, which is a feature the authors argue could complement screening tests such as the FOBT which show less sensitivity for right hand sided CRC compared to left sided.[69 92] No studies identified in this review applied machine learning approaches to model development. This approach warrants further investigation and comparison to more conventional methods such as logistic regression.

The studies which applied the model as a test to estimate diagnostic accuracy parameters in future studies should consider the patient flow and timing of the prediction model application. This can be achieved by avoiding a case-control study design which has been shown to bias test accuracy measures. Prospective study designs where all eligible patients are subjected to the index test (in this case the risk prediction model) as well as the reference standard compared to the FIT used on its own (as a comparator of standard care). This will determine the added benefit of using a risk adjusted approach to screening.

5.0 CONCLUSIONS AND RECOMMENDATIONS

Although it could not be formally tested in a meta-analysis due to the underlying heterogeneity in studies (outcomes, statistical methodology, populations, FITs and reported outcomes), there is some evidence to suggest that including additional factors with the FIT result can improve model performance and test accuracy for those studies using FIT only as a comparator to risk adjusted FIT. Improvement is seen in studies using

both lab based predictors as well as those including routine demographic data or richer questionnaire data. The latter predictors require no additional expensive lab based testing. Due to the limited number of studies, further evidence would be required to validate this improvement.

The majority of models which were identified in this review have several limitations in terms of statistical methodology for model development or with the sample size/population used for analysis. None of the models were considered currently usable in the target population in a screening based context. With the recent publication of the TRIPOD guidelines, future risk prediction modelling studies should consider these criteria when developing a new model and reporting the study in order to improve the quality of the study, provide reproducibility of the model development process and so the model can be validated externally or applied in an impact study. The model equation needs to be provided for this purpose including the intercept or a parameterisation of the baseline hazard if applying a survival analysis model.

Future research should focus on considering routine data which does not require additional data collection and is more likely to be complete to determine whether these predictors improve model performance and test accuracy. In addition, age and sex were factors consistently included in the final models and these should be considered in future model development studies along with consistent predictors identified in systematic reviews of risk prediction models which do not include the FIT. Machine learning methods have been used for similar studies and this could be a potential avenue to explore alongside more traditional statistical methods.

6.0 REFERENCES

1. Hewitson P, Glasziou P, Irwig L, Towler B, Watson E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. The Cochrane database of systematic reviews. 2007(1):Cd001216.
2. Allison JE, Fraser CG, Halloran SP, Young GP. Population screening for colorectal cancer means getting FIT: the past, present, and future of colorectal cancer screening using the fecal immunochemical test for hemoglobin (FIT). Gut and liver. 2014;8(2):117-30.
3. Benson VS, Patnick J, Davies AK, Nadel MR, Smith RA, Atkin WS. Colorectal cancer screening: a comparison of 35 initiatives in 17 countries. International journal of cancer Journal international du cancer. 2008;122(6):1357-67.
4. Halloran SP, Launoy G, Zappa M. European guidelines for quality assurance in colorectal cancer screening and diagnosis. First Edition--Faecal occult blood testing. Endoscopy. 2012;44 Suppl 3:Se65-87.
5. Launois R, Le Moine JG, Uzzan B, Fiestas Navarrete LI, Benamouzig R. Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. European journal of gastroenterology & hepatology. 2014;26(9):978-89.
6. van Rossum LG, van Rijn AF, Laheij RJ, van Oijen MG, Fockens P, van Krieken HH, et al. Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. Gastroenterology. 2008;135(1):82-90.
7. Moss S, Mathews C, Day TJ, Smith S, Halloran SP. A faecal immunochemical test for haemoglobin (FIT) markedly increased participation in a colorectal cancer screening pilot in England. Poster session presented at: Third NAEDI research conference March 26-27 2015. 2015.
8. Digby J, Fraser CG, Carey FA, McDonald PJ, Strachan JA, Diament RH, et al. Faecal haemoglobin concentration is related to severity of colorectal neoplasia. Journal of clinical pathology. 2013;66(5):415-9.
9. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. International journal of cancer Journal international du cancer. 2014;135(5):1203-12.
10. Garcia M, Mila N, Binefa G, Benito L, Gonzalo N, Moreno V. Fecal hemoglobin concentration as a measure of risk to tailor colorectal cancer screening: are we there yet? European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2015;24(4):321-7.
11. Imperiale TF, Glowinski EA, Lin-Cooper C, Ransohoff DF. Tailoring colorectal cancer screening by considering risk of advanced proximal neoplasia. American Journal of Medicine. 2012;125(12):1181-7.
12. Ma GK, Ladabaum U. Personalizing colorectal cancer screening: a systematic review of models to predict risk of colorectal neoplasia. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2014;12(10):1624-34.e1.
13. McDonald PJ, Strachan JA, Digby J, Steele RJ, Fraser CG. Faecal haemoglobin concentrations by gender and age: implications for population-based screening for colorectal cancer. Clinical chemistry and laboratory medicine : CCLM / FESCC. 2012;50(5):935-40.
14. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Colorectal cancer risk factors in the detection of advanced adenoma and colorectal cancer. Cancer epidemiology. 2013;37(3):278-83.

15. Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. The American journal of gastroenterology. 2009;104(3):739-50.
16. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
17. Gonzalez-Pons M, Cruz-Correa M. Colorectal Cancer Biomarkers: Where Are We Now? BioMed Research International. 2015;2015:149014.
18. Houlston RS. COGENT (COlorectal cancer GENeTics) revisited. Mutagenesis. 2012;27(2):143-51.
19. Tomlinson IPM, Dunlop M, Campbell H, Zanke B, Gallinger S, Hudson T, et al. COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. Br J Cancer. 2010;102(2):455-.
20. Tao S, Hoffmeister M, Brenner H. Development and validation of a scoring system to identify individuals at high risk for advanced colorectal neoplasms who should undergo colonoscopy screening. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2014;12(3):478-85.
21. Kaminski MF, Polkowski M, Kraszewska E, Rupinski M, Butruk E, Regula J. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. Gut. 2014;63(7):1112-9.
22. Spell DW, Jones DV, Harper WF, David Bessman J. The value of a complete blood count in predicting cancer of the colon. Cancer Detection and Prevention. 2004;28(1):37-42.
23. Brenner H, Haug U, Hundt S. Sex differences in performance of fecal occult blood testing. The American journal of gastroenterology. 2010;105(11):2457-64.
24. Alvarez-Urturi C, Andreu M, Hernandez C, Perez-Riquelme F, Carballo F, Ono A, et al. Impact of age- and gender-specific cut-off values for the fecal immunochemical test for hemoglobin in colorectal cancer screening. Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver. 2016;48(5):542-51.
25. Arana-Arri E, Idigoras I, Uranga B, Perez R, Irurzun A, Gutierrez-Ibarluzea I, et al. Population-based colorectal cancer screening programmes using a faecal immunochemical test: should faecal haemoglobin cut-offs differ by age and sex? BMC cancer. 2017;17(1):577.
26. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Medicine. 2013;10(2):e1001380.
27. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med. 2014;11(10):e1001744.
28. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ (Clinical research ed). 2017;356:i6460.
29. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Medicine. 2013;10(2):e1001381.
30. Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPDM-aMg. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. PLOS Medicine. 2015;12(10):e1001886.

31. Fraser CG, Rubeca T, Rapi S, Chen LS, Chen HH. Faecal haemoglobin concentrations vary with sex and age, but data are not transferable across geography for colorectal cancer screening. Clinical chemistry and laboratory medicine : CCLM / FESCC. 2014;52(8):1211-6.
32. Fraser CG, Auge JM. Faecal haemoglobin concentrations do vary across geography as well as with age and sex: ramifications for colorectal cancer screening. Clinical chemistry and laboratory medicine : CCLM / FESCC. 2015;53(9):e235-7.
33. Symonds EL, Osborne JM, Cole SR, Bampton PA, Fraser RJ, Young GP. Factors affecting faecal immunochemical test positive rates: demographic, pathological, behavioural and environmental variables. Journal of medical screening. 2015.
34. Kapidzic A, van der Meulen MP, Hol L, van Roon AH, Looman CW, Lansdorp-Vogelaar I, et al. Gender Differences in Fecal Immunochemical Test Performance for Early Detection of Colorectal Neoplasia. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2015;13(8):1464-71.e4.
35. UK National Screening Committee. The UK NSC recommendation on Bowel Cancer screening in adults 2016 [Available from: http://legacy.screening.nhs.uk/bowelcancer.
36. Moss S, Mathews C. NHS Bowel Cancer Screening Programmes: Evaluation of pilot of Faecal Immunochemical Test : Final report. National Screening Committee Website: Centre for Cancer Prevention, Wolfson Institute, Queen Mary University of London (QMUL); 2015.
37. van Veldhuizen H, Heijnen M-L, Lansdorp-Vogelaar I. Adjustment to the implementation of the colorectal cancer screening programme in 2014 and 2015. National Institute for Public Health and the Environment Ministry of Health, Welfare and Sport [Internet]. 2014 4th August 2015. Available from: http://www.rivm.nl/en/Documents_and_publications/Professional_Serviceable/Protocols/Disease_Prevention_and_Healthcare/Adjustment_to_the_implementation_of_the_colorectal_cancer_screening_programme_in_2014_and_2015.
38. World Endoscopy Organisation. 'FIT for Screening' Report. 6th Meeting of the Expert Working Group (EWG) Vienna. 2014.
39. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer prevention research (Philadelphia, Pa). 2016;9(1):13-26.
40. Williams TG, Cubiella J, Griffin SJ, Walter FM, Usher-Smith JA. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC gastroenterology. 2016;16(1):63.
41. Lee JK, Liles EG, Bent S, Levin TR, Corley DA. Accuracy of Fecal Immunochemical Tests for Colorectal CancerSystematic Review and Meta-analysis. Annals of Internal Medicine. 2014;160(3):171-81.
42. Shah R, Jones E, Vidart V, Kuppen PJ, Conti JA, Francis NK. Biomarkers for early detection of colorectal cancer and polyps: systematic review. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2014;23(9):1712-28.
43. Cochrane Colloquium Vienna, editor PROBAST: a risk of bias tool for prediction modelling studies. Cochrane Colloquium; 2015; Vienna.
44. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.
45. Dretzke J, Riley RD, Lordkipanidze M, Jowett S, O'Donnell J, Ensor J, et al. The prognostic utility of tests of platelet function for the detection of 'aspirin resistance' in patients with established cardiovascular or cerebrovascular disease: a systematic review

and economic evaluation. Health technology assessment (Winchester, England). 2015;19(37):1-366.
46. Pace N, Carlisle J, Eberhart L, Kranke P, Trivella M, Lee A, et al. Prediction models for the risk of postoperative nausea and vomiting (Protocol). Cochrane Database of Systematic Reviews 2014. 2014(9).
47. The Cochrane Collaboration. Cochrane Methods Prognosis 2015 [Available from: http://prognosismethods.cochrane.org/.
48. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ (Clinical research ed). 2009;339:b2700.
49. Carroll MR, Seaman HE, Halloran SP. Tests and investigations for colorectal cancer screening. Clinical biochemistry. 2014;47(10-11):921-39.
50. Halloran S, Launoy G, Zappa M. Faecal Occult Blood Testing. 2010 [cited 11th November 2014]. In: European Guidelines for Quality Assurance in Colorectal Cancer Screening and Diagnosis - First Edition [Internet]. Luxembourg: Publications Office of the European Union First. [cited 11th November 2014]; [103-44]. Available from: http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=ND3210390.
51. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
52. de Vet H, Eisinga A, Riphagen I, Aertgeerts B, Pewsner D. Searching for Studies. In: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [updated September 2008]. The Cochrane Collaboration. 2008.
53. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. Journal of the American Medical Informatics Association : JAMIA. 2001;8(4):391-7.
54. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons KG. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PloS one. 2012;7(2):e32844.
55. Barrows GH, Burton RM, Jarrett DD, Russell GG, Alford MD, Songster CL. Immunochemical detection of human blood in feces. American journal of clinical pathology. 1978;69(3):342-6.
56. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2006;174(4):469-76.
57. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med. 2003;138(1):W1-12.
58. Fraser CG, Halloran SP, Allison JE, Young GP. Making colorectal cancer screening FITTER for purpose with quantitative faecal immunochemical tests for haemoglobin (FIT). Clinical chemistry and laboratory medicine : CCLM / FESCC. 2013;51(11):2065-7.
59. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280-6.
60. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). Journal of the National Cancer Institute. 2005;97(16):1180-4.
61. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015;13:1.

62. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986;7(3):177-88.
63. Debray TP, Moons KG, editors. Systematic reviews of prognostic studies: a meta-analytical approach. Cochrane Prognostic Methods Group (PMG) Workshop. 2015 Cochrane Colloquium; 2015; Messe Congress Center, Vienna Austria.
64. Fraser CG, Allison JE, Halloran SP, Young GP. A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. Journal of the National Cancer Institute. 2012;104(11):810-4.
65. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology. 2005;58(10):982-90.
66. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Statistics in medicine. 2001;20(19):2865-84.
67. Boursi B, Mamtani R, Hwang WT, Haynes K, Yang YX. A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results. Digestive diseases and sciences. 2016;61(7):2076-86.
68. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer medicine. 2017;6(10):2453-60.
69. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. Journal of the American Medical Informatics Association : JAMIA. 2016;23(5):879-90.
70. Omata F, Shintani A, Isozaki M, Masuda K, Fujita Y, Fukui T. Diagnostic performance of quantitative fecal immunochemical test and multivariate prediction model for colorectal neoplasms in asymptomatic individuals. European journal of gastroenterology & hepatology. 2011;23(11):1036-41.
71. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014;147(3):628-+.
72. Kim BC, Joo J, Chang HJ, Yeo HY, Yoo BC, Park B, et al. A predictive model combining fecal Calgranulin B and fecal occult blood tests can improve the diagnosis of colorectal cancer. PloS one [Internet]. 2014; 9(9 // () *National Cancer Center* // () *National Cancer Center*). Available from: http://onlinelibrary.wiley.com/o/cochrane/clcentral/articles/537/CN-01014537/frame.html http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0106182&type=printable.
73. Tao S, Haug U, Kuhn K, Brenner H. Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. British Journal of Cancer. 2012;106(8):1424-30.
74. Wieten E, Grobbee EJ, Hansen BE, Bruno MJ, Kuipers EJ, Lansdorp-Vogelaar I, et al. Positive predictive value increases with age in a FIT-based colorectal cancer screening program. Gastroenterology. 2015;1):S760.
75. Karl J, Wild N, Tacke M, Andres H, Garczarek U, Rollinger W, et al. Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. Clinical Gastroenterology & Hepatology. 2008;6(10):1122-8.
76. National Institute for Health and Care Excellence. Quantitative faecal immunochemical tests to guide referral for colorectal cancer in primary care 2017 [Available from: https://www.nice.org.uk/guidance/dg30.

77. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006;332(7549):1080.
78. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.
79. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996;49(12):1373-9.
80. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine. 2016;35(2):214-26.
81. Pencina MJ, D' Agostino RB, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in medicine. 2008;27(2):157-72.
82. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 2010;21(1):128-38.
83. Hodgson LE, Sarnowski A, Roderick PJ, Dimitrov BD, Venn RM, Forni LG. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. BMJ open. 2017;7(9):e016591.
84. Ensor J, Riley RD, Moore D, Snell KI, Bayliss S, Fitzmaurice D. Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. BMJ open. 2016;6(5):e011190.
85. Jung YS, Park CH, Kim NH, Park JH, Park DI, Sohn CI. Clinical risk stratification model for advanced colorectal neoplasia in persons with negative fecal immunochemical test results. PloS one. 2018;13(1):e0191125.
86. Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, et al. A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. Gastrointestinal Endoscopy. 2015;81(3):719-27.
87. Yeoh KG, Ho KY, Chiu HM, Zhu F, Ching JY, Wu DC, et al. The Asia-Pacific Colorectal Screening score: a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects. Gut. 2011;60(9):1236-41.
88. Rodriguez-Alonso L, Rodriguez-Moranta F, Ruiz-Cerulla A, Lobaton T, Arajol C, Binefa G, et al. An urgent referral strategy for symptomatic patients with suspected colorectal cancer based on a quantitative immunochemical faecal occult blood test. Digestive and Liver Disease. 2015;47(9):797-804.
89. Murchie B, Tandon K, Hakim S, Shah K, O'Rourke C, Castro FJ. A New Scoring System to Predict the Risk for High-risk Adenoma and Comparison of Existing Risk Calculators. Journal of Clinical Gastroenterology. 2017;51(4):345-51.
90. van Vugt HA, Roobol MJ, Venderbos LD, Joosten-van Zwanenburg E, Essink-Bot ML, Steyerberg EW, et al. Informed decision making on PSA testing for the detection of prostate cancer: an evaluation of a leaflet with risk indicator. European journal of cancer (Oxford, England : 1990). 2010;46(3):669-77.
91. Edwards AG, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. The Cochrane database of systematic reviews. 2013(2):Cd001865.
92. Haug U, Kuntz KM, Knudsen AB, Hundt S, Brenner H. Sensitivity of immunochemical faecal occult blood testing for detecting left- vs right-sided colorectal neoplasia. Br J Cancer. 2011;104(11):1779-85.

7.0 APPENDICES

Appendix 1: Search Strategies

A.1.1 Medline (Ovid) Search Strategy 24/02/2016

#	Searches	Results
1	Validat$.mp. or Predict$.ti. or Rule$.mp. or (Predict$ and (Outcome$ or Risk$ or Model$)).mp. or ((History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$) and (Predict$ or Model$ or Decision$ or Identif$ or Prognos$)).mp. or (Decision$.mp. and ((Model$ or Clinical$).mp. or exp Logistic Models/)) or (Prognostic and (History or Variable$ or Criteria or Scor$ or Characteristic$ or Finding$ or Factor$ or Model$)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]	3,006,838
2	"Stratification".mp. or exp "ROC Curve"/ or "Discrimination".mp. or "Discriminate".mp. or "c-statistic".mp. or "c statistic".mp. or "Area under the curve".mp. or "AUC".mp. or "Calibration".mp. or "Indices".mp. or "Algorithm".mp. or "Multivariable".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]	527240
3	exp "Sensitivity and Specificity"/ or exp "Predictive Value of Tests"/ or sensitiv$.ti,ab. or specifici$.ti,ab. or accurac$.ti,ab. or "detection rate$".ti,ab. or "likelihood ratio$".ti,ab. or "false positive$".ti,ab. or "false negative$".ti,ab. or "predictive value$".ti,ab.	1641487
4	1 or 2 or 3	4382693
5	exp Colorectal Neoplasms/	161083
6	CRC.ti,ab.	13466
7	((cancer$ or neoplas$ or tumor$ or tumour$ or adenoma$ or adeno?carcinom$ or adenocarcinoma$ or polyp$) adj3 (colorectal$ or colon or rectal or rectum or bowel)).ti,ab.	128146
8	exp "Adenomatous Polyps"/	6999
9	"advanced adenoma".ti,ab.	321
10	5 or 6 or 7 or 8 or 9	189583
11	exp Occult Blood/ or exp Immunochemistry/ or ("fecal immunochemical" or "faecal immunochemical" or "fecal occult blood" or "faecal occult blood" or "FIT" or fobt$ or fob$ or ifobt or qnFIT or qlFIT or QTFIT or immunochemi$).ti,ab. or "OC-Micro".mp. or "OC Micro".mp. or "OC-Sensor".mp. or "OC Sensor".mp. or "OC-Hemodia".mp. or "OC Hemodia".mp. or "OC-	363615

	Light".mp. or "OC Light".mp. or HemeSelect.mp. or FlexSure.mp. or MagStream.mp. or "Instant-view".mp. or "Instant View".mp. or Hemoccult.mp. or Immocare.mp. or Monohaem.mp. or Hemosure.mp. or Ocultech.mp. or Quickvue.mp. or Clearview.mp. or Hemoquant.mp. or "Hema screen".mp. or "Hema-screen".mp. or Innovacon.mp. or Aimstep.mp. or Magstream.mp. or Immudia.mp.	
12	exp Early Detection of Cancer/ or exp mass screening/ or exp population surveillance/ or screen$.ti,ab. or (early adj3 detect$).ti,ab. or test$.ti,ab.	2626867
13	4 and 10 and 11 and 12	3037
14	limit 12 to (humans and yr="1978 -Current")	3009

A.1.2 Embase classic + Embase search strategy 24/02/2016 via OVID

#	Searches	Results
1	predict$.ti,ab. or validat$.ti,ab. or index.ti,ab. or model.ti,ab. or scor$.ti,ab. or exp "statistical model"/	4,415,041
2	"Stratification".mp. or exp "receiver operating characteristic"/ or "Discrimination".mp. or "Discriminate".mp. or "c-statistic".mp. or "c statistic".mp. or "Area under the curve".mp. or "AUC".mp. or "Calibration".mp. or "Indices".mp. or "Algorithm".mp. or "Multivariable".mp.	937201
3	exp "sensitivity and specificity"/ or exp "predictive value"/ or sensitiv$.ti,ab. or specifici$.ti,ab. or accurac$.ti,ab. or "detection rate$".ti,ab. or "likelihood ratio$".ti,ab. or "false positive$".ti,ab. or "false negative$".ti,ab. or "predictive value$".ti,ab.	2062939
4	1 or 2 or 3	6259870
5	exp colorectal tumour/ or exp colorectal cancer/ or exp colorectal adenoma/ or exp colorectal carcinoma/	132960
6	CRC.ti,ab.	26574
7	((cancer$ or neoplas$ or tumor$ or tumour$ or adenoma$ or adeno?carcinom$ or adenocarcinoma$ or polyp$) adj3 (colorectal$ or colon or rectal or rectum or bowel)).ti,ab.	200970
8	exp "adenomatous polyp"/	7975
9	"advanced adenoma".ti,ab.	837

10	5 or 6 or 7 or 8 or 9	243748
11	exp occult blood test/ or exp immunochemistry/ or ("fecal immunochemical" or "faecal immunochemical" or "fecal occult blood" or "faecal occult blood" or "FIT" or fobt$ or fob$ or ifobt or qnFIT or qlFIT or QTFIT or immunochemi$).ti,ab. or "OC-Micro".mp. or "OC Micro".mp. or "OC-Sensor".mp. or "OC Sensor".mp. or "OC-Hemodia".mp. or "OC Hemodia".mp. or "OC-Light".mp. or "OC Light".mp. or HemeSelect.mp. or FlexSure.mp. or MagStream.mp. or "Instant-view".mp. or "Instant View".mp. or Hemoccult.mp. or Immocare.mp. or Monohaem.mp. or Hemosure.mp. or Ocultech.mp. or Quickvue.mp. or Clearview.mp. or Hemoquant.mp. or "Hema screen".mp. or "Hema-screen".mp. or Innovacon.mp. or Aimstep.mp. or Magstream.mp. or Immudia.mp.	822038
12	exp screening/ or exp screening test/ or exp mass screening/ or exp population surveillance/ or screen$.ti,ab. or (early adj3 detect$).ti,ab. or test$.ti,ab.	4212903
13	4 and 10 and 11 and 12	3725
14	limit 12 to (humans and yr="1978 -Current")	3272

A.1.3 Cochrane Wiley Search 01/03/2016

#	Searches	Results
1	predict*:ti,ab or validat*:ti,ab or index:ti,ab or model:ti,ab or scor*:ti,ab	212929
2	MeSH descriptor: [Models, Statistical] explode all trees	15259
3	MeSH descriptor: [Risk Assessment] explode all trees	8672
4	stratification or discrimination or discriminate or "c-statistic" or "c statistic" or "area under the curve" or "AUC" or calibration or indices or algorithm or multivariable	42164
5	MeSH descriptor: [ROC Curve] explode all trees	1288
6	sensitiv*:ti,ab or specifici*:ti,ab or accurac*:ti,ab or "detection rate*":ti,ab or "likelihood ratio*":ti,ab or "false positive*":ti,ab or "false negative*":ti,ab or "predictive value*":ti,ab	48337
7	MeSH descriptor: [Sensitivity and Specificity] explode all trees	17833

8	MeSH descriptor: [Predictive Value of Tests] explode all trees	7035
9	(#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8)	282324
10	MeSH descriptor: [Colorectal Neoplasms] explode all trees	6053
11	CRC	1334
12	((cancer* or neoplasm* or tumor* or tumour* or adenoma* or adeno?carcinom* or adenocarcinoma* or polyp*) near/3 (colorectal* or colon or rectal or rectum or bowel))	11210
13	MeSH descriptor: [Adenomatous Polyps] explode all trees	175
14	advanced adenoma	69
15	(#10 or #11 or #12 or #13 or #14)	11865
16	MeSH descriptor: [Occult Blood] explode all trees	480
17	MeSH descriptor: [Immunochemistry] explode all trees	1282
18	fecal immunochemical:ti,ab or "faecal immunochemical":ti,ab or "fecal occult blood":ti,ab or "faecal occult blood":ti,ab or "FIT":ti,ab or fobt*:ti,ab or fob*:ti,ab or ifobt:ti,ab or qnFIT:ti,ab or qlFIT:ti,ab or QTFIT:ti,ab or immunochemi*:ti,ab	3807
19	OC-Micro or "OC Micro" or "OC-Sensor" or "OC Sensor" or "OC-Hemodia" or "OC Hemodia" or "OC-Light" or "OC Light" or HemeSelect or FlexSure or MagStream or "Instant-view" or "Instant View" or Hemoccult or Immocare or Monohaem or Hemosure or Ocultech or Quickvue or Clearview or Hemoquant or "Hema screen" or "Hema-screen" or Innovacon or Aimstep or Magstream or Immudia	151
20	(#16 or #17 or #18 or #19)	5295
21	MeSH descriptor: [Mass Screening] explode all trees	5443
22	MeSH descriptor: [Early Detection of Cancer] explode all trees	814
23	MeSH descriptor: [Population Surveillance] explode all trees	709
24	screen*:ti,ab or (early near/3 detect*):ti,ab or test*:ti,ab	178913
25	(#21 or #22 or #23 or #24)	180201
26	(#9 and #15 and #20 and #25)	321
27	(#9 and #15 and #20 and #25) * Publication Year from 1978 to 2016	**321**

A.1.4 Web of Science Core Collection minus Arts and Humanities 01/03/2016

#	Searches	Results
1	(TS=(predict* OR validat* OR index OR scor* OR (model* SAME statistical) OR (model* SAME (predict* OR decision*)) OR "risk assessment" OR stratification OR discrimination OR discriminate OR "c-statistic" OR "c statistic" OR "area under the curve" OR "AUC" OR calibration OR indices OR algorithm OR multivariable OR "ROC curve" OR "receiver operating characteristic"))	5,888,571
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
2	(TS=(sensitiv* OR specifici* OR accurac* OR "detection rate*" OR "likelihood ratio*" OR "false positive*" OR "false negative*" OR "predictive value*")) *AND* **LANGUAGE:** (English)	2,519,996
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
3	#2 OR #1	7,578,356
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
4	(TS=((cancer* OR neoplasm* OR tumor* OR tumour* OR adenoma* OR adeno?carcinom* OR adenocarcinoma* OR polyp*) NEAR/3 (colorectal* OR colon OR rectal OR rectum OR bowel))) *AND* **LANGUAGE:** (English)	207,090
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
5	(TS=("advanced adenoma" OR "adenomatous polyps" OR CRC))	20,473
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
6	#5 OR #4	211,507
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
7	(TS=("occult blood test" OR immunochemistry OR "fecal immunochemical" OR "faecal immunochemical" OR "fecal occult blood" OR "faecal occult blood" OR "FIT" OR fobt* OR fob* OR ifobt OR qnFIT OR qlFIT OR QTFIT OR immunochemi* OR "OC-Micro" OR "OC Micro" OR "OC-Sensor" OR "OC Sensor" OR "OC-Hemodia" OR "OC Hemodia" OR "OC-Light" OR "OC Light" OR HemeSelect OR FlexSure OR MagStream OR "Instant-view" OR "Instant View" OR Hemoccult OR	248,953

	Immocare OR Monohaem OR Hemosure OR Ocultech OR Quickvue OR Clearview OR Hemoquant OR "Hema screen" OR "Hema-screen" OR Innovacon OR Aimstep OR Magstream OR Immudia)) *AND* **LANGUAGE:** (English)	
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
8	(TS=(screen* or (early NEAR/3 detect*) or test* or surveillance)) *AND* **LANGUAGE:** (English)	4,369,037
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	
9	#8 AND #7 AND #6 AND #3	**1,978**
	Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, ESCI Timespan=1978-2016	

Appendix 2: Data Extraction Form based on the CHARMS Checklist.

Domain	**Key items to extract**
1. Study Characteristics/Design	- Author - Year - Describe the study design, whether cohort, case control, diagnostic accuracy study etc - Country - Setting
2. Source of Data	- Source of data (e.g. cohort, case-control, randomized trial participants, or registry data)
3. Participants	- Participant eligibility and recruitment method (e.g. consecutive participants, location, number of centres, setting, inclusion and exclusion criteria) - Participant description (e.g. females aged between 60-74 participating in a screening programme) - Study dates - Mean/median age and ethnicity
4. Outcomes to be predicted	- Definition and method for measurement of outcome - Was the same outcome definition (and method for measurement) used in all patients? - Type of outcome (single or combined endpoints) - Was the outcome assessed without knowledge of the candidate predictors (i.e. blinded)? - Were candidate predictors part of the outcome (e.g. in panel or consensus diagnosis)? - Diagnosis at colonoscopy or follow up? - Cancer detection rate if applicable
5. Candidate predictors/risk factors (or index tests)	- Number and type of predictors (e.g. demographics, patient history, physical examination, additional testing, disease characteristics) - Definition and method for measurement of candidate predictors - Timing of predictor measurement (e.g. at patient presentation, at diagnosis, at treatment initiation) - Were predictors assessed blinded for outcome, and for each other (if relevant)? - Handling of predictors in the modeling (e.g. continuous, linear, non-linear transformations or categorized) -Corresponding adjusted and unadjusted odd ratios etc with measures of uncertainty (95% confidence interval and P values)
6. Sample Size	- Number of participants and number of outcomes/events - Number of outcomes/events in relation to the number of candidate predictors (events per variable)
7. Missing data	- Number of participants with any missing value (include predictor and outcomes) - Number of predictors with missing data for each predictor - Handling of missing data (e.g. complete case analysis, imputation or other methods)

8. Model Development	- Modelling method (e.g. logistic, survival, neural networks or machine learning techniques) - Modelling assumptions satisfied - Method for selection of predictors for inclusion in multivariable modeling (e.g. all candidate predictors, pre-selection based on unadjusted association with the outcome) - Method for selection of predictors during multivariable modeling (e.g. full model approach, backward or forward selection) and criteria used (e.g. p-value, Akaike Information Criterion) - Shrinkage of predictor weights or regression coefficients (e.g. no shrinkage, uniform shrinkage, penalized estimation) - The corresponding relative risks, odds ratios or hazard ratios (adjusted and unadjusted) with estimates of uncertainty for the predictors included (95% confidence intervals/P values)
9. Model Performance	- Calibration (calibration plot, calibration slope, Hosmer-Lemeshow test) and Discrimination (C-statistic, D-statistic, log-rank) measures with confidence intervals. - Classification measures (e.g. sensitivity, specificity, predictive values, net reclassification improvement, integrated discrimination improvement) and whether a priori cut points were used
10. Model Evaluation	- Method used for testing model performance: development dataset only (random split of data, resampling methods, e.g. bootstrap or cross-validation, none) or separate external validation (e.g. temporal, geographical, different setting, different investigators) - In case of poor validation, whether model was adjusted or updated (e.g. intercept recalibrated, predictor effects adjusted, or new predictors added)
11. Risk Stratification	- **How is risk stratification combined with the FIT?** What method has been used to combine risk stratification with the FIT? (E.g. Modellling: unvariable analysis, multivariable analysis (including logistic regression and time-to-event analysis), nomograms, artificial neural networks, decision trees) Or another method? **- How is risk assessed?** Is risk assessed using a score, using risk tiers/categories, as a probability etc? **- When is risk assessment carried out?** Before/after the index and reference tests? (recall bias)
12. Diagnostic Accuracy Considerations	-Target disease definition -Type and brand of FIT (qualitative FIT or quantitative FIT, OC-Sensor, OC-Micro etc) -Reference Standard and rationale (e.g. Colonoscopy or follow up?) - Number of FIT samples -Prevalence of Colorectal cancer or advanced adenomas in the study - Number of participants enrolled for the study - Number of participants for whom results are available - Methods and timing of index test - Methods and timing of reference standard
13. Test accuracy	- 2 x 2 data -Sensitivity -Specificity -AUC ROC -Cutpoint and reason for cutpoint -(if presented) True positives, false positives, true negatives and false negatives -Positive and negative predictive values -Positive and negative diagnostic likelihood ratios -The CRC and advanced adenoma detection rate (representing the benefits of screening) -(If included) NNTScope (representing the harms of screening) -Number of missing, indeterminate, intermediate and uninterpretable test results.
14. Results	- Final and other multivariable models (e.g. basic, extended, simplified) presented, including predictor weights or regression coefficients, intercept, baseline survival, model performance measures (with standard errors or confidence intervals) - Any alternative presentation of the final prediction models e.g. sum score, nomogram, score chart, predictions for specific risk subgroups with

	performance - Comparison of the distribution of predictors (including missing data) for development and validation datasets
15. Interpretation and Discussion	- Interpretation of presented models (confirmatory, i.e. model useful for practice versus exploratory i.e. more research needed) - The clinical applicability of study findings - Comparison with other studies, discussion of generalizability, strengths and limitations

Appendix 3: Exclusion of Studies

#	Full Text	
	Article	**Reason for Exclusion**
1	Aniwan, S., et al. (2015). "A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants." Gastrointestinal endoscopy 81(3): 719-727.	Model The study just looks at prevalence (and sensitivity/specificity) in 4 different groups based on the combination of the FIT result and APCS score. They are not combining this information in a risk prediction model for individualised risk prediction.
2	Cai, S. R., et al. (2011). "Performance of a colorectal cancer screening protocol in an economically and medically underserved population." Cancer Prevention Research 4(10): 1572-1579.	Model Does not combine the questionnaire and FIT into a prediction model for individualised predictions.
3	Calistri, D., et al. (2010). "Fecal DNA for noninvasive diagnosis of colorectal cancer in immunochemical fecal occult blood test-positive individuals." Cancer Epidemiology, Biomarkers & Prevention 19(10): 2647-2654.	Model Discussed with 3rd reviewer (NP) - the model underlying the nomogram is unclear. It is not strictly a model development study. Model performance parameters not reported.
4	Cha, J. M., et al. (2012). "First-degree relatives of colorectal cancer patients are likely to show advanced colorectal neoplasia despite a negative fecal immunochemical test." Digestion **86**(4): 283-287.	Model FIT result is not included as dependent variable in the logistic regression. The outcome is defined as 'colorectal cancer despite a negative FIT result'.
5	Chang, L. C., et al. (2014) Metabolic syndrome and smoking may justify earlier colorectal cancer screening in men. Gastrointestinal endoscopy 79, 961-969 DOI: 10.1016/j.gie.2013.11.035	Model Model does not give individual risk, no details about probability, just looks at adjusted odds ratios. Exclude based on population and performance measures.
6	Chen, H. S. and S. M. Sheen-Chen (2002). "Influence of age and gender on surveillance for colorectal tumors in low-risk asymptomatic population." Anticancer Research 22(1A): 399-403.	Model Not a risk prediction model development study FOBT is probably guaiac
7	Chen, L. S., et al. (2011). "Baseline faecal occult blood concentration as a predictor of incident colorectal neoplasia: longitudinal follow-up of a Taiwanese population-based colorectal cancer screening cohort." Lancet Oncology 12(6): 551-558.	Model Not producing individualised risk just looking at associations in a multivariable model. Timing considered in the model does not match review question.
8	Chiang, T. H., et al. (2014). "Difference in performance of fecal immunochemical tests with the same hemoglobin cutoff concentration in a nationwide colorectal cancer screening program." Gastroenterology 147(6): 1317-1326.	Model Compares the outcomes between two different types of FIT. Looking at the difference in performance between two FITs 'Adjusting for' other factors Not an individualised risk prediction model
9	Garcia, M., et al. (2015). "Fecal hemoglobin concentration as a measure of risk to tailor colorectal cancer screening: are we there yet?" European Journal of Cancer Prevention 24(4): 321-327.	Model Logistic regression model just looked at associations not individualised risk
10	Imperiale, T. F., et al. (2014). "Multitarget stool DNA testing for colorectal-cancer screening." N Engl J Med **370**(14): 1287-1297.	Model (creating a panel test) The algorithm used for this study consists of three parts to obtain an overall score; a logistic score which combines all of the DNA and Hb results, a Sum of Scores which incorporates the Logistic Score and individual marker scores and a Composite score (Sum of Scores is subjected to an

		exponential equation). The multi-target stool DNA test provides a dichotomous positive or negative result. No other information is provided. The logistic score is not used as the final risk prediction model, the result of this is used to calculate a Sum of Scores then a Composite Score – and this Composite Score is then designated positive/negative. In addition, the hemoglobin component of the multi-target sDNA Test is not a commercially available FIT. The main aim of the study is to determine the performance characteristics of the DNA test in the detection of CRC.
11	KIM, N. H., KWON, M. J., KIM, H. Y., LEE, T., JEONG, S. H., PARK, D. I., CHOI, K. & JUNG, Y. S. 2016. Fecal hemoglobin concentration is useful for risk stratification of advanced colorectal neoplasia. *Dig Liver Dis,* 48**,** 667-72.	Model Model investigates risk in terms of odds ratios but not individualised risk. This needs to go one step further as it is not a risk prediction model.
12	Lidgard, G. P., et al. (2013). "Clinical performance of an automated stool DNA assay for detection of colorectal neoplasia." Clin Gastroenterol Hepatol 11(10): 1313-1318.	Model The aim of the study is to optimise an automated sDNA assay and evaluate its clinical performance. Part of this process is combining DNA markers and hemoglobin component in a logistic regression algorithm – but determining the overall output from this algorithm involves two other components as described in the supplementary material in the study by Imperiale *et al*. 2014. 'The individual results are used in combination to develop a logistic regression algorithm which generated a dichotomous patient result' Test – The hemoglobin component of the multi-target sDNA test is not a commercially available FIT.
13	Park, M. J., et al. (2012). "A comparison of qualitative and quantitative fecal immunochemical tests in the Korean national colorectal cancer screening program." Scandinavian Journal of Gastroenterology 47(4): 461-466.	Model Combines FIT type in the model not the FIT result. Not developing a model for individualised risk prediction just looking at positivity rates.
14	Qin, M., et al. (2015). "Risk factors for colorectal neoplasms based on colonoscopy and pathological diagnoses of Chinese citizens: a multicenter, case-control study." International Journal of Colorectal Disease 30(3): 353-361.	Model Model does not give individualised predictions. It just looks at ORs/associations. Population is patients from hospitals. Does not state which FOBT is used.
15	Rengucci, C., et al. (2014). "Improved stool DNA integrity method for early colorectal cancer diagnosis." Cancer Epidemiology Biomarkers and Prevention 23(11): 2553-2560.	Model Not strictly a risk prediction model development study. Does not report any model performance/test parameters for the combination. Tested whether FL-DNA assay could improve diagnostic accuracy. Could argue Bayesian method. This was discussed with a third author (NP) who agreed this is not strictly a model development study and the methods are not clear.
16	Wild, N., et al. (2010). "A combination of serum markers for the early detection of colorectal cancer." Clinical Cancer Research 16(24): 6111-6121.	Model Does combine FIT with serum markers but just looks at diagnostic accuracy. The aim is not to develop a risk prediction model with individualised risk prediction.
17	Wong, M. C. S., et al. (2014). "Should prior FIT results be incorporated as an additional variable to estimate risk of colorectal neoplasia? A prospective study of 5,813 screening colonoscopies." PLoS ONE	Model Not individualised risk prediction model, just investigates associations.

18	Zheng, S., et al. (2003). "Cluster randomization trial of sequence mass screening for colorectal cancer." Diseases of the Colon and Rectum 46(1): 51-58.	Model FOBT result does not contribute to ADV score but is used to decide on which risk group. Not individualised risk prediction combining FIT with risk factor detail in a statistical model Flexible sigmoidoscopy used as diagnostic tool only small numbers of colonoscopy for a specific protocol. The associated paper which assesses the risk assessment is in Chinese.
19	Parente, F., et al. (2012). "A combination of faecal tests for the detection of colon cancer: a new strategy for an appropriate selection of referrals to colonoscopy? A prospective multicentre Italian study." European Journal of Gastroenterology & Hepatology 24(10): 1145-1152.	Population - Population is all symptomatic and symptoms are not included in the model There is also no model it uses a 'if one positive' rule between the tests
20	Rodriguez-Alonso, L., et al. (2015). "An urgent referral strategy for symptomatic patients with suspected colorectal cancer based on a quantitative immunochemical faecal occult blood test." Digestive and Liver Disease 47(9): 797-804.	Population Model for primary care use. Population is primary care for urgent referral.
21	Sanders, A. D., et al. (2013). "A novel pathway for investigation of colorectal symptoms with colonoscopy or computed tomography colonography." New Zealand Medical Journal 126(1382): 45-57.	Population Not known whether the FOBT is FIT or not - probably guaiac as its not specified. Symptomatic population only in primary care This study assesses a whole pathway, the scoring system is just a part of it.
22	Brazer, S. R., et al. (1991). "USING ORDINAL LOGISTIC-REGRESSION TO ESTIMATE THE LIKELIHOOD OF COLORECTAL NEOPLASIA." Journal of Clinical Epidemiology 44(11): 1263-1270.	Screening Test Guaiac based test referenced in the text also reference 22 in the text relates to a guaiac test
23	Griffiths, E. K. and D. V. Schapira (1991). "Serum ferritin and stool occult blood and colon cancer screening." Cancer Detection & Prevention 15(4): 303-305.	Screening Test Seracult guaiac test
24	Kaminski, M. F., et al. (2014). "A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy." Gut 63(7): 1112-1119.	Screening Test Does not combine FIT within the prediction model.
25	Sequist, T. D., et al. (2011). "Electronic Patient Messages to Promote Colorectal Cancer Screening A Randomized Controlled Trial." Archives of Internal Medicine **171**(7): 636-641.	Screening Test The personalised model /scoring system behind/underlying this study does not combine FIT. Not a risk prediction modelling study design. Study looks at uptake, not test or model performance May have applicability issues as they look at patients overdue for colorectal cancer screening who are at higher risk
26	Bosch, L. J., et al. (2012). "DNA methylation of phosphatase and actin regulator 3 detects colorectal cancer in stool and complements FIT." Cancer Prevention Research 5(3): 464-472.	Study Design Not a model development study- just combines the positive result of one test with the positive result of another. Referral subjects – uses just stool samples Does not produce individualised risk predictions
27	Brenner, H., et al. (2014). "Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy." Gastroenterology **146**(3): 709-717.	Study Design The authors investigated specific associations of CRC risk with previous colonoscopy conducted for various indications which does not match the review question. Guaiac based test The population had various indications, including primary screening, surveillance after a preceding colonoscopy, follow-up of positive FOBT result, or specific symptoms.
28	Cubiella, J., et al. (2014). "Diagnostic accuracy of	Study Design

	fecal immunochemical test in average- and familial-risk colorectal cancer screening." United European Gastroenterology Journal **2**(6): 522-529.	Issues with study design.
29	Huang, W., et al. (2014). "Cost-effectiveness of colorectal cancer screening protocols in urban Chinese populations." PLoS ONE [Electronic Resource] **9**(10): e109150.	Study Design Study design looks at cost effectiveness, not a prediction modelling study. FOBT is probably guaiac too as it is not explicitly specified.
	Abstracts	
	Article	**Reason for Exclusion**
1	Sassatelli, R., et al. (2014). "Repeated fit screening: The influence of subject's characteristics and screening history on the positive predictive value and neoplasia yield." Gastroenterology 1): S410-S411.	Model Does not mention a statistical model. PPV and DR are the outcomes Discrimination or calibration not mentioned
2	Ardizzoia, A., et al. (2011). "A combination of fecal tests for the detection of colon cancer: A new strategy for appropriate prioritization of referrals to colonscopy-A prospective Italian study." Journal of Clinical Oncology. Conference: ASCO Annual Meeting 29(15 SUPPL. 1).	Population Includes those with abdominal symptoms and it is uncertain whether/how they combine the predictors/tests. It is uncertain whether individualised predictions are determined.
3	Guery, E., et al. (2015). "Performance of a blood methylated mutitarget DNA test (colohybritest) for the diagnosis of colorectal (CRC) cancer: A transversal study on 878 individuals." Gastroenterology 1): S745.	Population The population is considered at high risk by including symptomatic patients as well as those with a positive gFOBT. Blood test is adjusted for age, positive FIT, symptoms family history. Does not mention individualised risk prediction.
4	Boursi, S. B., et al. (2015). "Impact of a risk model based on routine lab results on colorectal cancer screening in average risk population." Journal of Clinical Oncology. Conference 33(15 SUPPL. 1).	Screening Test FIT is not included in the model
5	Castiglione, G., et al. (1986). "THE PREDICTIVE VALUE OF HEMOCCULT AND AN ANAMNESTIC QUESTIONNAIRE FOR THE EARLY DETECTION OF COLORECTAL-CANCER IN SELF SELECTED PATIENTS." Digestive Diseases and Sciences 31(10): S462-S462.	Screening Test Hemoccult is Guaiac test. Page 460, abstract 1837
6	Imperiale, T., et al. (2013). "Risk for advanced colorectal neoplasia in asymptomatic adults is effectively stratified by phenotypic features." American Journal of Gastroenterology 108: S646.	Screening Test Does not mention FIT in the model. Associated with a full text: Imperiale, T. F., et al. (2015). "Derivation and Validation of a Scoring System to Stratify Risk for Advanced Colorectal Neoplasia in Asymptomatic Adults: A Cross-sectional Study." Ann Intern Med **163**(5): 339-346.
7	Imperiale, T., et al. (2014). "A scoring system for predicting the risk of advanced proximal neoplasia in asymptomatic adults." American Journal of Gastroenterology 109: S600.	Screening Test Does not mention FIT in the model. Associated with a full text
8	* Ladabaum, U., et al. (2015). "Potential effectiveness and cost-effectiveness of tailoring screening to predicted colorectal cancer risk." Gastroenterology 1): S1.	Study Design Author was contacted for further information. The study was a hypothetical modelling exercise which did not apply to the review.
9	Aniwan, S., et al. (2013). "Clinical risk score and fecal immunochemical testing are helpful to prioritize the patients for colon cancer screening by colonoscopy." Gastroenterology 1): S578.	Full Text Associated Although different numbers probably a follow on study with different risk groups defined too. They are not combining the information in a risk prediction model for individualised risk prediction.
10	Auge, J. M. (2014). "Practical experience of the fecal hemoglobin immunochemical test in a colorectal cancer screening program." Anticancer Research 34 (10): 5821.	Full Text Associated The risk component is associated with a full text
11	Auge, J. M. (2014). "Fit and colonoscopy. Competition and cooperation." Tumor Biology 35: S17.	Full Text Associated The risk component is associated with a full text
12	Kim, N., et al. (2015). "Fecal hemoglobin concentration is useful for risk stratification of advanced colorectal neoplasia." Journal of Gastroenterology and Hepatology (Australia) 30: 94-95.	Full Text Associated

13	Lidgard, G. P., et al. (2012). "An optimised multi-marker stool test for colorectal cancer screening: Initial clinical appraisal." Gastroenterology 1): S770.	Full Text Associated
14	Lidgard, G. P., et al. (2012). "An optimised molecular stool test for colorectal cancer screening: Evaluation of an automated analytic platform and logistic algorithm." Cancer Prevention Research. Conference: 11th Annual AACR International Conference on Frontiers in Cancer Prevention Research Anaheim, CA United States. Conference Start 5(11 SUPPL. 1).	Full Text Associated
15	Wieten, E., et al. (2015). "Positive predictive value increases with age in a FIT-based colorectal cancer screening program." Gastroenterology 1): S760.	Full Text Associated which is included as a final text.
16	Wild, N., et al. (2010). "Early detection of colorectal cancer applying a combination of serum markers." Cancer Research. Conference: 101st Annual Meeting of the American Association for Cancer Research, AACR 70(8 SUPPL. 1).	Full Text Associated
17	*Bosch, L. J. W., et al. (2015). "Advanced neoplasia detection in colorectal cancer screening using multiple stool DNA markers and haemoglobin." Journal of Pathology 237: S14.	Author could not supply further detail due to impending publication of results. The FIT was however combined in a statistical model with the sDNA panel.

*contacted author for additional information

Appendix 4: Assessment of Methodological Quality

A.4.1 QUADAS-2 Tailored Tool: Risk of bias and applicability judgements

QUADAS-2: Systematic review of risk prediction models combining the FIT result for Colorectal Cancer screening

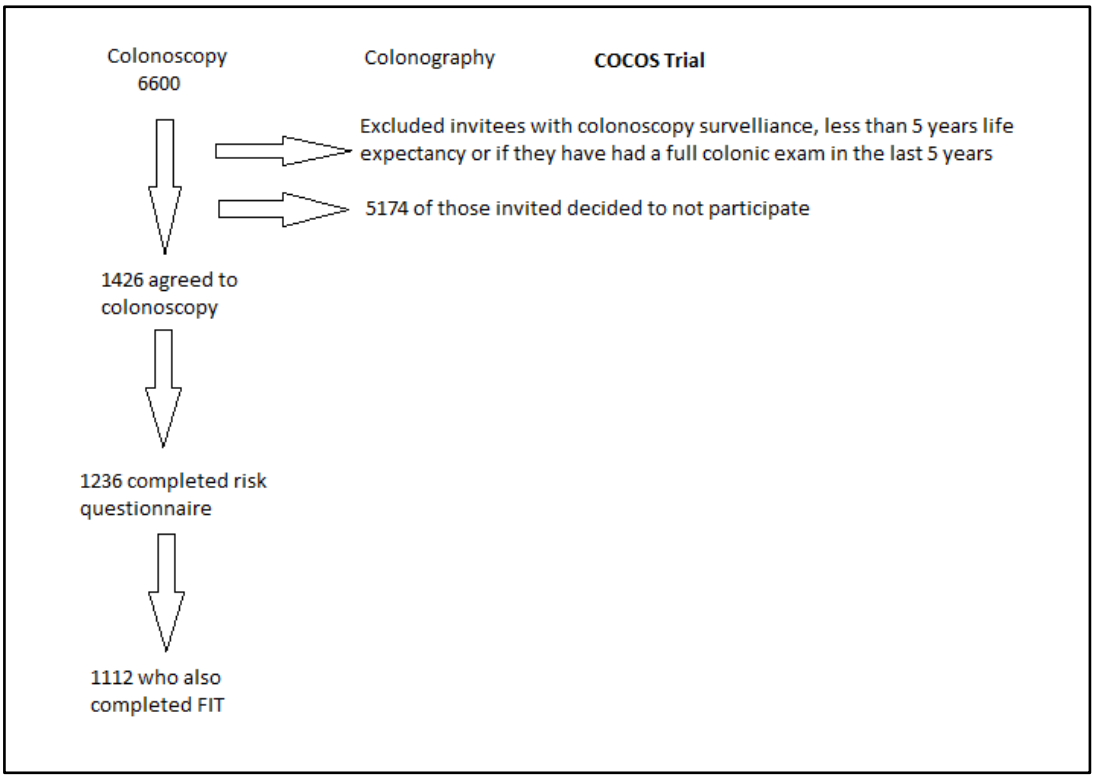
First author surname and year of publication: ***Stegeman 2014***

Name of first reviewer: Jennifer Cooper *Name of second reviewer: Karoline Freeman*

Phase 1: State the review question:

Patients (setting, intended use of index test, presentation, prior testing): *Both men and women aged 40-75 years representative of an average risk screening population (mean/average age needs to be over 40 years so participants over 18 can be included/10% outside screening range is acceptable)*
Index test(s): *Where the FIT has been combined with other risk factors/predictors in a risk prediction model/risk scoring system (e.g. risk model, score or clinical decision rule) This may or may not be compared to using FIT alone*
Reference standard and target condition: *Where diagnostic accuracy parameters are included, the reference standard can be either; colonoscopy or at least two years of follow up using clinical records (cancer registries, GP records etc). Small numbers of other diagnostic procedures acceptable as some individuals may not be suitable for colonoscopy e.g. CT/flexible sigmoidoscopy). The target condition is advanced neoplasia which includes both colorectal cancer and advanced adenomas as well as colorectal polyps.*

Phase 2: Draw a flow diagram for the primary study



Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgements regarding bias and applicability. For each section, one 'NO' = High Risk of bias.

DOMAIN 1: PATIENT SELECTION **A. Risk of Bias**	
Describe methods of patient selection: Data collected from randomized screening trial in the Netherlands (COCOS study) 6600 asymptomatic men and women 50-75 years of age randomly selected and invited for colonoscopy All those who underwent colonoscopy were asked to complete risk questionnaire and a FIT Exclusions – those who have had colonic exam in the last 5 years, those in colonoscopy surveillance programme, those with a life expectancy of less than 5 years.	
+ Was a consecutive or random sample of patients enrolled?	Yes/No/Unclear
• Yes: If a study has enrolled all consecutive or a random sample of eligible patients with suspected disease • No: If a study makes inappropriate exclusions or studies which enrol patients with known disease status. • Unclear: when insufficient data are reported to decide.	
+ Was a case-control design avoided?	Yes/No/Unclear
• Yes: if a case–control design was avoided (nested case-control study have less bias) • No: if a case–control design was used and data on predictors collected after the diagnosis was made. Case-control studies create spectrum bias which can overestimate the accuracy of a test. Recall bias can also be a problem but if a study uses data collected on predictors before the diagnosis was made (i.e. GP records) then this can be minimised. • Unclear: when insufficient data are reported to decide.	
+ Did the study avoid inappropriate exclusions?	Yes/No/Unclear
• Yes: if a study avoided inappropriate exclusions. (Acceptable exclusions could be comorbidities, if a patient had been previously diagnosed with CRC or if they were on a colonoscopy surveillance programme as an example) • No: if a study made inappropriate exclusions. • Unclear: when insufficient data are reported to decide.	
Could the selection of patients have introduced bias? If all signaling questions for a domain are answered 'yes' there is a low risk of bias, if any signaling question is answered 'no' this indicates a high level of bias.	**RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability	
Describe included patients (prior testing, presentation, intended use of index test and setting): 6600 asymptomatic men and women 50-75 years of age randomly selected and invited for colonoscopy as part of the COCOS randomized screening trial in the Netherlands. 1426 agreed to undergo colonoscopy, 1236/1426 completed the questionnaire and 1112 also completed the FIT. Intended use of the risk prediction model combining FIT is pre-selection for colonoscopy in a screening programme setting.	
Is there concern that the included patients do not match the review question?	**CONCERN: LOW/HIGH/UNCLEAR**
• High concern: if the study population cannot be considered an adult average risk screening population (men and women aged 40-75), if individuals are known to have high risk genetic	

syndromes such as familial adenomatous polyposis (FAP) or Lynch syndrome. Hospitalised patients or symptomatic patients only would also be at higher risk and would indicate high concern of applicability issues. • Low concern: if the study population can be considered an adult average risk screening population (men and women aged 40-75) • Unclear: when insufficient data are reported to decide.

DOMAIN 2: INDEX TEST(S) **If more than one index test was used, please complete for each test.** The FIT and risk prediction model that has been applied as a test need to be considered in this section. **A. Risk of Bias**	
1 - Describe the index test and how it was conducted and interpreted: 2 - Describe how the risk factor information is collected (before or after index/reference standard, by questionnaire/routine electronic data/chart review etc) 3 – Impact study considers the model outcome (probability/risk score); model development considers predictor information including the FIT. The OC-Sensor FIT test was conducted before colonoscopy was carried out (one sample). A threshold of 50 ng/ml was used to compare the sensitivity of the risk model at a specificity of 93%. Risk factor information was collected via a questionnaire which was handed out to the participants in the waiting room before colonoscopy (therefore patients have not changed responses according to the results – recall bias).	
+ Were the index test results (predictors including FIT/probability or risk score) interpreted without knowledge of the results of the reference standard? (Test review bias) Yes as FIT and questionnaire completed before colonoscopy	Yes/No/Unclear
• Yes: If the analysts are blinded (masked) to the results of the reference investigation and other clinical information (knowledge of additional clinical information is clinical review bias) (colonoscopy/two year follow up/small numbers of alternative investigations) • No: if the FIT result or risk model was interpreted with knowledge of the reference standard (colonoscopy/two year follow up) • Unclear: when insufficient data are reported to decide.	
+ Was risk information obtained around the time of the FIT result but before the reference standard (or previously recorded on clinical records)? I.e. avoiding retrospective data collection? (Recall bias) This includes if more than one lab test incorporated.	Yes/No/Unclear
• Yes: If retrospective data collection is avoided or obtained after the reference standard • No: If retrospective risk information collected, or any lab tests performed after the reference standard • Unclear: when insufficient data are reported to decide	
+ Is the brand/type of FIT used described?	Yes/No/Unclear
• Yes: if the study mentioned the brand/type of FIT i.e. whether qualitative or quantitative. (Ideally name of specimen collection device and supplier given) • No: if the study does not describe the brand/type of FIT • Unclear: when insufficient data are reported to decide.	
+If a threshold was used, was it pre-specified? Used 10 µg Hb/g of faeces as this was the anticipated cut off for the Dutch screening programme i.e. it was not reverse engineered to give positive results.	Yes/No/Unclear

• Yes: if the study indicates what threshold of the FIT is used for a positive test result or sets the sensitivity/referral rate/probability threshold the same if comparing a risk scoring system incorporating the FIT with the FIT alone. The threshold considered needs to be within the range used by a screening programme; 20 µg Hb/g of faeces is commonly used as a cutoff for the OC-Sensor test as an example. Published studies have considered thresholds around 10-180 µg Hb/g. • No: if the study does not pre-specify a threshold or selects a threshold solely to maximise performance. • Unclear: when insufficient data are reported to decide.	
+ Details of faecal collection method (sampling technique and number of samples if applicable) provided. One sample but the sampling technique is not reported.	Yes/No/Unclear
• Yes: if the study indicates the sampling technique • No: if the study does not indicate the sampling technique • Unclear: when insufficient data are reported to decide.	
+ Was the FIT return time up to 10 days only or was the test repeated if past this date? (This could give false negative results)[1]	Yes/No/Unclear
• Yes: if the study indicates that the FIT result is returned within 10 days of the test/excludes those after this time period or repeats the test. • No: if the study includes the FIT result if returned within 10 days of the test/includes those results after this time period or doesn't repeat the test • Unclear: when insufficient data are reported to decide.	
+ Time and storage of collection devices from specimen collection to analysis, including time and temperature (median and range) indicated?	Yes/No/Unclear
• Yes: if the study indicates the time and storage of collection devices from specimen collection to analysis • No: if the study does not include the time and storage of collection devices from specimen collection to analysis • Unclear: when insufficient data are reported to decide.	
Could the conduct or interpretation of the index test have introduced bias? The first 4 signaling questions are the most important for risk prediction models. If these 4 are Yes then low risk of bias. The first four signaling questions are Yes.	**RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability **Is there concern that the index test, its conduct, or interpretation differ from the review question?** Quantitative FIT OC Sensor and risk questionnaire (used to develop the risk prediction model) are performed before colonoscopy.	**CONCERN: LOW/HIGH/UNCLEAR**

[1] van Roon, A. H., et al. (2012). "Are fecal immunochemical test characteristics influenced by sample return time? A population-based colorectal cancer screening trial." Am J Gastroenterol **107**(1): 99-107.

- High Concern: If there are variations in test technology, execution or interpretation which would affect diagnostic accuracy
- Low Concern: If test methods are those specified in the review question (i.e. qualitative or quantitative FIT)
- Unclear: when insufficient data are reported to decide.

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

Colonoscopies were performed using the standard quality aspects defined by the American Society for Gastrointestinal Endoscopy. Colonoscopy was performed after the risk questionnaire and FIT. The most advanced lesion per patient was used. Advanced neoplasia was defined as at least one CRC or advanced adenoma: adenoma of 10 mm or larger, ≥25% villous histology or high grade dysplasia. CRC and advanced adenoma were reported separately. All included participants had a colonoscopy.

+ Were reference standard results interpreted without knowledge of the results of the index test? (Diagnostic review bias)	Yes/No/Unclear

- Yes: if colonoscopy is carried out without knowledge of the FIT result/other clinical information (clinical review bias) – more difficult to meet this criteria in screening studies as usually positive results are referred on for colonoscopy.
- No: if colonoscopy is carried out with knowledge of the FIT result/risk model result
- Unclear: if insufficient information is provided to decide.

+ Is the reference standard likely to correctly identify CRC?	Yes/No/Unclear

- Yes: if the reference standard was colonoscopy (considered the gold standard) or two year follow up. Small numbers of other diagnostic tests such as CT/Flexible Sigmoidoscopy are acceptable if the patient is not suitable for colonoscopy – these need to be noted as a large proportion will bias results.
- No: if the reference standard was not was colonoscopy (considered the gold standard) or two year follow up.
- Unclear: when insufficient data are reported to decide.

+ Has the study avoided introducing partial or differential verification bias? Partial verification bias is particularly an issue for screening studies whereby only those with positive results are referred for diagnostic testing.[2]	Yes/No/Unclear

- Yes: If all study subjects who have had an index test have also received the reference standard (colonoscopy) or if there is at least two years of follow up (or verification bias is corrected using mathematical correction methods). Ideally all patients should have the same reference standard (e.g. just colonoscopy) but this is often not the case in screening studies.
- No: If not all study subjects who have had an index test have also received the reference standard (colonoscopy), or if there is not at least two years follow up (or mathematical correction has not been applied). Bias will also be present if a different reference standard is used like flexible sigmoidoscopy in different groups of people.
- Unclear: when insufficient data are reported to decide.

Could the reference standard, its conduct, or its interpretation have introduced bias? **RISK: LOW/HIGH/UNCLEAR**

[2] de Groot, J. A. H., et al. (2011). "Verification problems in diagnostic accuracy studies: consequences and solutions." BMJ **343**.

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question?

CONCERN: LOW/HIGH/UNCLEAR

- High: If the target condition that the reference standard defines differs from the target condition (colorectal cancer, advanced adenoma, colorectal polyps) in the review question.
- Low: if the target condition that the reference standard defines does not differ from the target condition (colorectal cancer, advanced adenoma, colorectal polyps) in the review question.
- Unclear: when insufficient data are reported to decide.

DOMAIN 4: FLOW AND TIMING

A. Risk of Bias

Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):

In total 6600 persons were invited for primary colonoscopy screening, of which 1426 (22%) agreed to undergo colonoscopy. In this group, 1236 (87%) individuals completed the questionnaire and 1112 (90%) of them also completed the FIT test.

+Was there an appropriate interval between index test (predictors and FIT/probability or risk score) and reference standard? FIT completed before the colonoscopy along with the questionnaire.	Yes/No/Unclear

- Yes: if there is an appropriate interval between FIT and colonoscopy. Abnormal tests are offered an appointment with a SSP clinic within 14 days and a colonoscopy appointment is made within two weeks (under 6-8 weeks could be considered a reasonable time interval for a chronic condition.[3]
- No: if there is not an appropriate interval between FIT and colonoscopy
- Unclear: when insufficient data are reported to decide.

+ Did all patients receive a reference standard? Out of the original people invited 1426 had colonoscopy. All patients who had a FIT would have had a colonoscopy.	Yes/No/Unclear

- Yes: if all patients received a reference standard (colonoscopy/two year follow up).
- No: if all patients did not receive a reference standard (colonoscopy/two year follow up).
- Unclear: when insufficient data are reported to decide.

+ Did all patients receive the same reference standard?	Yes/No/Unclear

- Yes: if the reference standard was the same (colonoscopy/two year follow up). A systematic review by Lee *et al.*[4] shows that the sensitivity estimates vary when using colonoscopy on FIT negative patients (0.71) compared with a two year follow up (0.87) so if there is a mix this

[3] Logan, R. F., et al. (2012). "Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests." Gut **61**(10): 1439-1446.

[4] Lee, J. K., et al. (2014). "Accuracy of Fecal Immunochemical Tests for Colorectal CancerSystematic Review and Meta-analysis." Annals of Internal Medicine **160**(3): 171-181.

needs to be considered in assessing risk of bias. • No: if the reference standard was not the same (colonoscopy/two year follow up). • Unclear: when insufficient data are reported to decide.
+ Were all patients included in the analysis? (all patients who were recruited into the study should be included in the analysis) — Yes/No/Unclear
• Yes: if all participants were included i.e. if 2 by 2 data is presented, the number of patients enrolled should not differ from the number included in the 2 by 2 tables. Those who drop out from having follow up tests may be at higher risk due to lifestyle factors, comorbidities, or if they are not appropriate for further diagnostic tests. • No: if participants were excluded. • Unclear: when insufficient data are reported to decide.
Could the patient flow have introduced bias? **RISK: LOW/HIGH/UNCLEAR** All those enrolled were not included in the final analysis. Not all participants completed both a FIT and questionnaire. Those who agreed to have a colonoscopy are more likely to be healthy – the healthy screenee effect which could potentially reduce sensitivity.

DOMAIN 5: ROLE OF SPONSOR **A. Risk of Bias**
+ Did the funding source/sponsor play no role in design of study, interpretation of results and publication? Yes/No/Unclear '…funded by The Netherlands Organization for Health Research and Development of the Dutch Ministry of Health… The sponsor was not involved in the study.'
Could the funding source have introduced bias? **RISK: LOW/HIGH/UNCLEAR** • High: If the FIT manufacturer as an example have played a role in the design of the study, interpretation of results and publication. • Low: if the funding source/sponsor have not played a role in the design of the study, interpretation of results or publication. • Unclear: when insufficient data are reported to decide.

A.4.2 PROBAST: Risk of bias and applicability judgements

(Prediction model study Risk of Bias Assessment Tool)

An early version of PROBAST was piloted but not reproduced in this thesis since the finalised tool will be published in 2018.

PROBAST assesses risk of bias and applicability of studies evaluating a multivariable diagnostic or prognostic prediction model. The tool is used for studies which develop a model or externally validate a model.

There are 5 key domains including: Participant Selection, Predictors, Outcome, Sample Size and Participant Flow and Statistical Analysis. Signalling questions are rated as 'Yes', 'Probably Yes', 'Probably No', 'No' or 'No Information'. Each domain is then subsequently rated as 'high', 'low' or 'unclear' risk of bias. An overall judgement is rated for each model along with the usability of the model. The first three domains listed above are also assessed for applicability to the review question.

Risk-adjusted Colorectal Cancer Screening Using the FIT and Routine Screening Data: Development of a Risk Prediction Model

Chapter based on the following published paper: Cooper, J. A., et al. (2018). "Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model." Br J Cancer **118**(2): 285-293.

ABSTRACT

Objectives: The fecal immunochemical test (FIT) has recently been recommended to replace the guaiac fecal occult blood test in the NHS Bowel Cancer Screening Programme. Increased uptake and positivity of the FIT will put added strain on limited colonoscopy services. There is evidence to suggest that combining risk indicators such as age, sex and other lifestyle factors with the FIT improves test accuracy and may help to guide more efficient colonoscopy use. This study aimed to develop a risk prediction model combining routinely available predictors from the Bowel Cancer Screening System with the FIT to determine whether model performance and test accuracy are improved in a representative sample of the English screening population.

Design: Data for this study were collected during the six-month FIT pilot study for the NHS Bowel Cancer Screening Programme. Of the people invited to complete a FIT kit, 27,066 individuals aged 59-75 adequately participated. Multivariable analysis used those with a positive FIT (≥20 μg/g) and with a diagnostic colonoscopy outcome (n=1810). Stepwise backwards elimination was used to build a logistic regression model using FIT result, age, sex and previous screening history. Model outcome was either cancer or advanced adenoma (high-risk adenoma or intermediate-risk adenoma) detected at colonoscopy. Model performance was assessed using discrimination and calibration and test accuracy was investigated using sensitivity, specificity and receiver operating characteristic (ROC) curves.

Results: Of the sample used for multivariable analysis (n=1810), 549 cancers and advanced adenomas were detected (30.3%). Discrimination improved from 0.628 with FIT only to 0.659 for the risk-adjusted model (p=0.01). The calibration of the risk-adjusted model was 0.898 using the Hosmer-Lemeshow statistic compared with 0.481 for FIT only. The sensitivity improved from 30.78% (FIT only) to 33.15% (risk-adjusted model) at similar specificity using a threshold of 160 μg/g which is the anticipated threshold for the NHS

Bowel Cancer Screening Programme. The risk-adjusted screening algorithm detected 13 more advanced adenomas and the same number of cancers compared to FIT-only at a threshold of 160 µg Hb/g faeces. The risk model mainly improved detection in men, but also more than halved the number of false positive results for women. When analysing the detection rates by subgroup, there was a reduction in the detection rate for female first time invitees. The detection rate for male previous non-responders more than doubled (16.98% to 37.11%).

Conclusions: Risk-adjusted screening using routinely available predictors on the BCSS enhanced both model performance and test accuracy. This could lead to more appropriate referrals whereby those at greatest risk are sent for a follow up colonoscopy. Further investigation is required relating to the greater cancer/advanced adenoma detection in males compared with females and the acceptability of this difference to the population and the screening programme. Future research should investigate additional predictors available from the Bowel Cancer Screening System, particularly relating to screening history, as well as other modelling approaches including machine learning methods.

1.0 INTRODUCTION

The previous chapter reported a systematic review of risk prediction models which combine the FIT result with other risk predictors for colorectal cancer screening referral decisions. There was some evidence to suggest that including additional factors with the FIT result can improve model performance and test accuracy when comparing FIT only to a risk based FIT model. Both lab based predictors and routine data (e.g. age, sex, BMI) lead to an improvement in discrimination as well as test accuracy. The advantages of using routine data only are that no further testing is required and it often has a high percentage of completeness. The Bowel Cancer Screening System (BCSS) used in the NHS stores detailed information for all participants invited to screening in an electronic health record format. This data source was investigated as a potential method to enhance the performance of the FIT.

In England, a 6-month comparative pilot study was initiated by the NHS BCSP in April 2014 to assess uptake and acceptability, as well as diagnostic performance of FIT.[1] The FIT pilot using the OC-Sensor/DIANA *Eiken Chemical Co., Tokyo, Japan* (FIT) found an improved uptake (66.4% versus 59.3%) and increased cancer detection rate using FIT compared to the gFOBT. Due to this increased uptake of the test and higher positivity, additional strain could be expected on colonoscopy services. The pilot results suggest an additional 290,000 people would be screened each year when replacing the current test with the FIT, leading to further referrals for colonoscopy.[1] Furthermore, those who participate in earlier screening rounds have been shown to be more likely to continue participating in later rounds.[2 3] One paper reports that England has had a 20% increase in colonoscopy capacity over the last 5 years, with a current rate of 360,000 examinations performed each year.[4] A range of thresholds are currently being investigated for the FIT between 150 and 180 µg Hb/g faeces to keep the referral rates similar to current colonoscopy service use.[1 5]

Another approach which could improve effective colonoscopy use, test accuracy and consequently health outcomes is risk based colorectal cancer screening.[1 6 7] A few studies have developed risk prediction models which combine the FIT concentration with other risk indicators for use in screening referral decisions.[7-13] For example, Stegeman *et al*[8] in the Netherlands combined the FIT with risk indicators obtained from a lifestyle questionnaire in a logistic regression model. Sex, age, calcium intake, smoking status, family history and the

FIT result were retained in the final model which had improved sensitivity (40% versus 32% for FIT alone) at a 93% specificity.[8] The Netherlands is currently in the process of implementing a pragmatic integrated trial of risk-adjusted FIT with outcome diagnostic yield.[14] Other studies have investigated combining blood based inflammatory markers with the FIT which improved discrimination from 0.683 to 0.729 (AUC ROC),[11] or integrating demographic characteristics to categorise risk into 16 different subgroups.[7] The Asia-Pacific Screening Scoring System which includes age, sex, smoking and family history was combined in an algorithm with FIT to triage participants for colonoscopy based on whether they were at low, medium or high risk.[15]

The studies described hitherto required additional testing or lifestyle questionnaires to obtain predictor information for the model. Sending additional documents such as questionnaires has been shown to significantly reduce screening uptake.[16] A more efficient approach would be to utilise screening data that are routinely available as an electronic record. This would reduce participant burden and enhance data completeness for use within a prediction model. As far as can be identified, the present study is the first to adopt this pragmatic approach in a population-based screening programme and to combine risk indicators with the FIT in an English screening population. Following the transition to the FIT for the NHS BCSP, application of this approach could improve test accuracy and enable more efficient use of limited and expensive colonoscopy resources.

The aim of this study was to develop a risk prediction model which integrates routinely available predictors from the NHS BCSS with the quantitative FIT result to determine whether model performance and test accuracy are improved in an English screening population.

2.0 METHODS

Since this study both develops a risk prediction model and assesses test accuracy, the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and Standards for Reporting of Diagnostic Accuracy Studies (STARD) statements have been followed for reporting.[17 18] In addition, Steyerberg's checklist (**Table 1**) for developing valid prediction models was considered for data management and analysis to ensure that an internally valid prediction model was produced.[19]

Step	Specific Issues
General Considerations	
Research question	Aim: predictors/prediction?
Intended application	Clinical practice/research?
Outcome	Clinically relevant?
Predictors	Reliable measurement? Comprehensiveness
Study Design	Retrospective/prospective? Cohort; case-control
Statistical Model	Appropriate for research question and type of outcome?
Sample Size	Sufficient for aim?
Seven Modelling Steps	
Data Inspection	Missing values
Coding of Predictors	Continuous predictors Combining categorical predictors Restrictions on candidate predictors
Model Specification	Appropriate selection of main effects? Assessment of assumptions (distributional, linearity, and additivity)
Model Estimation	Shrinkage included?
Model Performance	Appropriate statistical measures used? Clinical usefulness considered?
Model Validation	Internal validation, including model specification and estimation? External Validation?
Model Presentation	Format appropriate for audience?
Validity	
Internal: Overfitting	Sufficient attempts to limit and correct for overfitting?
External: Generalizability	Predictions valid for plausibly related populations?

Table 1: Steyerberg's checklist for developing valid prediction models.[19]

2.1 Study population and data source

The NHS BCSP performed a comparative study to determine the acceptability and accuracy of the FIT compared with the gFOBT.[1] The study involved two out of the five regional screening hubs in England; (i) the Midlands and North West Hub and (ii) the Southern Hub. Between 7th April and 10th October 2014, 1,126,087 individuals were invited to complete a gFOBT, with 667,945 adequately screened (i.e. those with a definitive positive or negative result), and 40,930 individuals were invited to complete a FIT (one out of 28 screening invitations), with 27,167 adequately screened. The pilot analysed data from participants aged 59 to 75 years old. The FIT pilot study is discussed in further detail elsewhere.[1]

This analysis is limited to complete cases (i.e. those with complete data records) and those who had a FIT result of 20 µg Hb/g faeces and above (n=1810) as this was the cutoff chosen for test positivity during the pilot and ensured participants had a definitive diagnosis at colonoscopy. Twenty µg Hb/g faeces was the threshold chosen for test positivity for the pilot. Since this threshold was set low, it enabled different thresholds to be examined up to 180 µg Hb/g faeces and their corresponding effects on positivity and clinical outcomes. This approach also enables investigation of a risk-adjusted approach by comparing relative performance between FIT only and FIT combined with risk factors.

The data for the FIT pilot and this study were recorded on the BCSS which contains routine information on the screening pathway for participants. These data were anonymised and provided by the Health and Social Care Information Centre (HSCIC) – now NHS Digital - through the Office for Data Release (ODR). Data were extracted by NHS Digital on the 10th March 2016.

All participants who received an invitation during the study period were included (7 April 2014 to 10 October 2014). For the sample population analysed, FIT kits were distributed between 15th April 2014 and 19th November 2014. Completed kits were received at the lab between 22nd April 2014 and 5th March 2015 and examined between 25th April 2014 and 9th March 2015.

2.2 Ethical Approval

Ethical approval was obtained from the University of Warwick Biomedical and Scientific Research Ethics Committee (BSREC) (Reference Number REGO-2015-1575). The Bowel Cancer Screening Research Committee also approved the study protocol (ID152). The data were anonymised and provided by NHS Digital through the Office for Data Release (ODR1516_045). The approval letters along with the System Level Security Policy followed for this project are provided in **Appendix 1** and **Appendix 2** respectively.

2.3 Routinely available predictors

The routinely available predictors recorded on the BCSS which were investigated in this study were age, sex, Index of Multiple Deprivation (IMD) score and previous screening history (i.e. whether someone was a previous non-responder/responder to screening). Age at the start of the screening episode for the pilot was used for the analysis. Social Deprivation was measured using the IMD score, which is derived using the English Indices of Deprivation 2010 based on 38 separate indicators in seven domains.[20] The IMD is calculated for every Lower Super Output Area (LSOA) in England and can be ranked to give the IMD rank. There are 32,482 LSOA in the UK, which are small areas based on postcode and consisting of approximately 1500 people.[20] IMD score was supplied by NHS Digital based on the participant's postcode of residence.

2.4 FIT Concentration (Index test)

The OC-SENSOR FIT (Eiken Chemical Co. Ltd., Japan, supplied by Mast Diagnostics, UK) was used along with the OC-SENSOR Diana analyser. The FIT units were converted from ng Hb/ml buffer to µg Hb/g faeces as recommended by the World Endoscopy Organisation and experts, to aid comparison between studies.[21] The OC-SENSOR deposits 10mg of faeces into 2.0mL of buffer. Units in nanograms of haemoglobin per ml of buffer can be converted to µg haemoglobin per g faeces by multiplying the concentration by 2 and dividing by 10 for certain FITs (100ng haemoglobin per mL of buffer is equal to 20 µg Hb/g faeces).[21] FIT kits were sent by post for participants to complete at home and returned by mail to the screening hubs.

2.5 Colonoscopy (Diagnostic test)

Participants with a positive test result were offered a specialist screening practitioner appointment and referred for a colonoscopy assessment within 14 days of this appointment (alternative investigations were arranged if the colonoscopy was inappropriate or fails first time round e.g. CT scan or flexible sigmoidoscopy).[22] Colonoscopies were performed using the quality assurance guidelines for colonoscopy published by the NHS Cancer Screening Programmes.[23]

The NHS BCSS uses an algorithm to record the diagnosis of an individual based on the guidelines for colorectal cancer screening and surveillance (**Table 2**). A low risk adenoma is indicated by the presence of 1 or 2 small adenomas less than 1cm in diameter. An intermediate risk adenoma is indicated by the presence of 3 or 4 small adenomas or at least one which is ≥1cm in diameter. A high risk adenoma is indicated by the presence of ≥5 adenomas or ≥3 adenomas where at least one is ≥1cm in diameter.[24]

Adenoma Type	Definition
Low Risk Adenoma	The presence of 1 or 2 small adenomas which are less than 1cm in diameter
Intermediate Risk Adenoma	The presence of 3 or 4 small adenomas or at least one which is ≥1cm in diameter
High Risk Adenoma	The presence of ≥5 adenomas or ≥3 adenomas where at least one is ≥1cm in diameter

Table 2: Definition of adenomas used by the NHS Bowel Cancer Screening System based on the guidelines for colorectal cancer screening and surveillance.

2.6 Model Outcome

The binary model outcome was colorectal cancer or advanced adenoma detected at colonoscopy after a positive FIT referral. Advanced adenomas were those classified as either high risk or intermediate risk, since these have potential, if left untreated, to develop into bowel cancer, particularly as age increases.[25 26] Detection of these adenomas is a key health outcome of the NHS bowel cancer screening programme (NHSBCSP).[22] In addition, the NHS BCSP carries out surveillance of patients with high or intermediate risk adenomas which have been detected at previous screening rounds.[22 24]

For modelling purposes, colorectal cancer or advanced adenoma was coded as a single binary outcome variable (i.e. yes = 1 or no = 0). Where there was more than one diagnostic outcome recorded for an individual, the 'greatest risk' scenario was used giving one diagnostic outcome per individual.

2.7 Statistical analysis

All data were analysed in RStudio Version 0.99.903 (driven by R version 3.3.1) on a Windows 7 computer.[27] Additional packages were also loaded from The Comprehensive R Archive Network (CRAN; https://cran.r-project.org/).[28-34] Two models were tested using logistic regression, with a binary response variable of cancer status; (i) FIT concentration only as a predictor and (ii) FIT concentration and routine data. The R scripts used to develop and assess the performance of the models are provided in **Appendix 3**.

2.7.1 Model Development

Routinely available risk factors from the database were selected based on previous studies[7 8] and the information available from the data extract provided. Initial evaluation of routinely collected predictor variables and their association with colorectal cancer and advanced adenomas was undertaken using univariable logistic regression.

Typically, univariable screening of predictors which are over a pre-defined p-value is not recommended,[35] therefore when developing the risk model all variables were considered for inclusion. The risk-adjusted model was built by adding all the routinely available risk factors into a single multivariable logistic regression model and then using backwards elimination to remove non-significant variables with a p-value greater than 0.1 as

determined through likelihood ratio testing. Backwards elimination is generally preferred to stepwise forwards selection as the latter is more likely to exclude predictors involved in suppressor effects.[36] The z statistic was used to assess the significance of estimated model coefficients; this follows a normal distribution and indicates whether the coefficient is significantly different from 0.[36] All possible pairwise interactions were also investigated.

Models developed using this methodology often perform poorly when used on new data; i.e. different data from that which was used for model development.[37 38] Models are over-fitted to the particular data used for model development and thus do not generalise well to new settings.[37] One way to address this issue is to divide the data into training and validation datasets, the former being used for model building and the latter for model performance testing. When data are limited, a more efficient procedure is to use cross-validation[39] as an internal validation method. Cross-validation involves partitioning the data sample into distinct subsets, performing the analysis on one subset (training data), and validating the analysis on the other subset (validation data). Multiple runs of cross-validation are performed using random partitions of the data, which are then averaged over the runs. Ten-fold cross-validation was used during model development; the data were divided 10 times into 10 randomly selected subsets of approximately equal size and 90% of the data were used for model building (training data) and predictions were made on the remaining 10% of the data. Bootstrapping (10,000 replications) was used to construct confidence intervals for the model coefficients and to estimate bias. The confidence intervals reflect our true confidence in the value calculated from the data alone.[38]

All continuous variables were kept as such (i.e. not dichotomised or categorised) as recommended by previous research and the TRIPOD guidelines (age, FIT concentration, IMD score).[37 40-42] Age was not formally significant in the final model but was retained as it was found to be a consistent significant predictor in the models identified in **Chapter 2** (along with sex).[7 8 43] The log of the FIT concentration was used for analysis as this was found to give improved model performance and better fitted linearity. Screening history was coded as a factor (either a previous responder, previous non-responder or first-time invitee). This was determined using two variables recorded on the BCSS; sequence number (EPISODE_SEQUENCE_NUMBER) and type of episode (PREVALENT_INCIDENT). Sex on the BCSS was treated as a factor with two levels: male or female.

2.7.2 Model Performance

Model performance was assessed using calibration and discrimination. Calibration assesses the agreement between the observed outcomes and the predictions.[44] This was determined using the Hosmer-Lemeshow statistic where a small p-value indicates that the model does not predict accurately.[45] The observations can be split into a different number of groups based on their predicted probabilities and this can affect the corresponding result of the Hosmer-Lemeshow statistic.[46] Therefore, group splits between 5 and 15 were investigated. A calibration plot of predicted risk versus observed risk was plotted for deciles of participants where points close to the line of equality indicates good calibration.[19]

Discrimination is the ability of a model or test to distinguish between those with and without the outcome (cancer status) or those at high risk versus those at low risk.[44] This was assessed using the c-statistic, which is also referred to the AUC (area under the curve) of the Receiver Operating Characteristic (ROC) curve.

The two models (risk-adjusted model and FIT only model) were compared using the likelihood ratio test (which quantifies changes in the model residual deviance). Overall model fit for the logistic regression models was assessed using pseudo R^2 measures including Nagelkerke's R^2. Ordinary Least Squares regression uses the standard R^2 measure to assess goodness of fit (how well the model fits the data) by comparing the null to the fitted model but there is not an equivalent measure for assessing logistic regression. Therefore, several pseudo R^2 measures have been proposed. These measures usually range from 0 to 1 with higher values indicating a better model fit. Cox and Snell's R^2 assesses the improvement of the full model over the intercept model. Since this statistic cannot reach the maximum value of one, Nagelkerke introduced an amendment so the possible value can extend to one.[47]

2.7.3 Test Accuracy of the Risk Model

The ROC curves were plotted for the risk-adjusted FIT model and FIT only to compare test accuracy across different thresholds. A ROC test was performed to analyse the difference between the AUC for both models using bootstrapping.

Two by two tables were then produced to determine the sensitivity and specificity for a threshold of 160µg Hb/g faeces and the equivalent risk threshold for the risk-adjusted model. This threshold was selected based on previous work from the FIT pilot,[1] as well as discussions regarding colonoscopy capacity by stakeholders. A threshold of 150 µg Hb/g faeces gave a similar positivity rate to the gFOBT and 180 µg Hb/g faeces a similar referral rate.[1 5] A threshold of between 150 µg Hb/g faeces and 180 µg Hb/g faeces will be adopted by the NHS Bowel Cancer Screening programme and adjusted as appropriate in order to ensure adequate colonoscopy capacity.

The predictiveness curve, which plots the distribution of risk, has been proposed by Pepe *et al.*[48] as an alternative plot which allows both the assessment of the fit of the model and the clinical utility when applied to the population. It is argued that the predictiveness curve can give additional information about risk-threshold which is not typically provided by the ROC curve. A histogram of risk probabilities can display similar information however this involves defining risk intervals/bins whereas the predictiveness curves shows the proportions of patients over risk thresholds.[48]

To plot a prediction curve, estimated risk for each individual is calculated from the logistic regression model and ordered from highest to lowest, this is plotted against risk percentile of the sample population. The predictiveness curve will be plotted for both the risk-adjusted model and the model for FIT only for comparison.

2.7.4 Additional predictors and their effect on FIT positivity

The effect of sample return time, ambient temperature, sex and age on FIT positivity at a threshold of 160 µg Hb/g faeces was investigated in unvariable logistic regression. The sample population for this analysis was those who had adequately participated with a definitive FIT result of positive or negative (n=27,066). At 160 µg Hb/g faeces, 26,614

participants had a negative result and 452 had a positive result and there were 14,305 females (52.85%) and 12,761 males.

Sample Return Time

Sample return time can affect FIT screening outcomes due to the degradation of haemoglobin present in the sample, leading to false negative results.[49] The manufacturer Eiken Chemical Co., Ltd state that the stability of Haemoglobin for the OC-SENSOR should last up to 14 days if the ambient temperature is between 2 and 10 ˚C or up to 7 days if stored at room temperature. The time taken from sample (F_SAMPLE_DATE) to laboratory receipt (DATE_KIT_LOGGED) was investigated using the corresponding dates recorded on the BCSS.

Ambient Temperature

Ambient temperature has been shown to affect the positivity rate of the FIT.[50-53] Ambient temperature that the FIT was exposed to once the sample had been completed was estimated by determining the median time from sample date (F_SAMPLE_DATE) to when the FIT was received at the lab (DATE_KIT_LOGGED) (2 days). This gives a day for the sample, day in the post and a day for receipt at the lab so it was assumed the FIT would be exposed to ambient temperature for 3 days. The daily maximum temperature across these days for each participant was then determined and this variable was named 'maximum mean temperature'. This is a similar method to that of Daley *et al.*[51] who investigated the effect of ambient temperature on positivity rate.

Temperature data were obtained from the UK Met Office, providing data for each weather station in the UK.[54] Data were downloaded as CSV files from http://datamarket.azure.com/dataset/datagovuk/metofficeweatheropendata. This dataset it provided to data.gov.uk by the UK Met Office and is hosted by Windows Azure Datamarket. The Church Lawford station (site code 3544) was the nearest station to the Midlands hub and Farnborough weather station (site code 3768) was the nearest weather station to the Southern hub. These stations were used as a regional estimate for every participant within the region.

2.8 Reproducing the Dataset

Data were provided by NHS Digital as separate dataframes: Subjects, Episodes, Testkits, Colonoscopy Assessment, Diagnostic and Polyps. These dataframes were cleaned and merged for the analysis as appropriate. A data schematic, and how the data were merged for analysis, is given in **Figure 1** below.

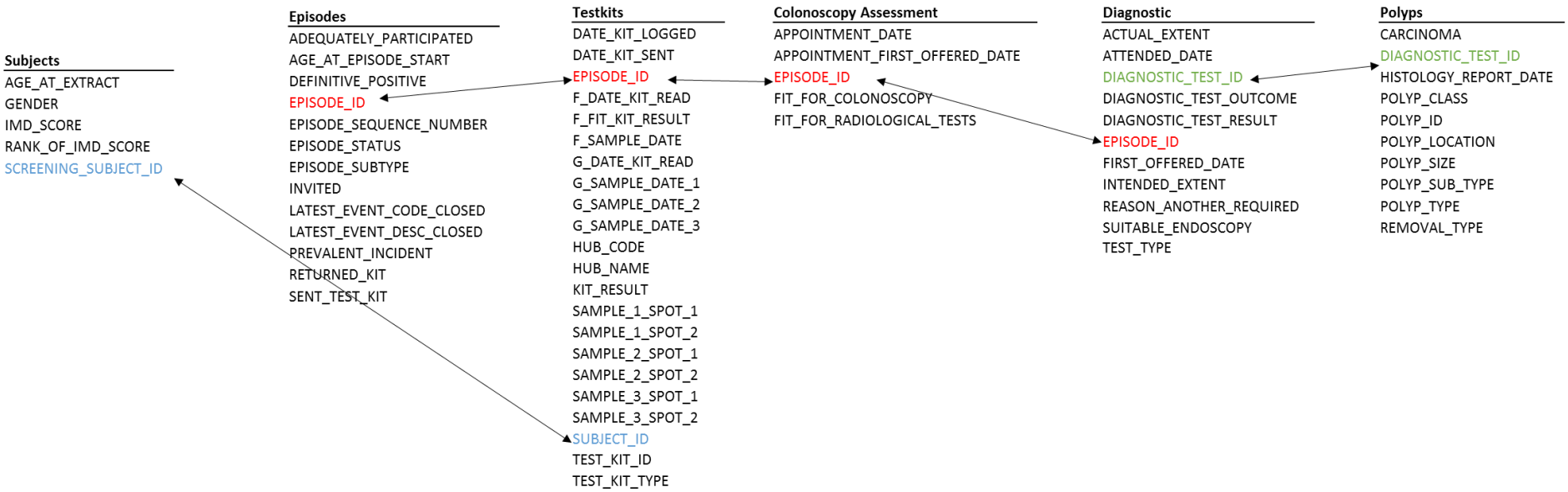


Figure 1: Data schematic and linkage across dataframes for the FIT pilot data extract provided from NHS Digital

The 'Subject' dataframe was merged with the 'Testkit' dataframe and subsetted to FIT tests only (n=40,930). The FIT kit results were then limited to normal/abnormal which removed the results which were classed as 'spoilt'. Individuals with multiple test kits were limited to the most recent result giving a one to one relationship for subject and test kit. The 'Episodes' dataframe was then merged with this dataset to obtain the LATEST_EVENT_DESC_CLOSED field which was used to determine the definitive outcome for an individual. This final dataset was then limited to those individuals who adequately participated (those who returned the FIT kit and have a definitive result) (n=27,066). To investigate the demographics of those with an abnormal test kit the dataset was limited to those with a FIT result of ≥20 µg Hb/g faeces (n=2117) this sample population was also used for univariable analysis. For multivariable analysis and model building, complete cases were used which limited the sample size to n=1810.

The LATEST_EVENT_DESC_CLOSED variable in combination with the 'Diagnostic' dataframes when needed was used to provide the definitive diagnosis. This field gives the latest status of that episode when it is closed. Where there was just a description and not a diagnostic outcome (e.g. Hand over into Symptomatic Care) the 'Colonoscopy Assessment', 'Diagnostic' and 'Polyps' dataframes were investigated for each individual to determine the diagnostic outcome. Where there was more than one diagnostic outcome recorded for an individual, the 'greatest risk' scenario was used. Where a diagnostic appointment was made and an individual did not attend, this was classified as 'Not attended' and where an appointment was cancelled the outcome was classified as 'Cancelled' (See **Appendix 4** to see how individuals were classified). The possible outcome categories for an individual were: Subject Discharge (Normal), Abnormal, Low-risk Adenoma, Normal (No Abnormalities Found), Intermediate-risk Adenoma, High-risk Adenoma, Cancer, Cancelled and Not Attended.

2.9 Estimation of Test Accuracy Measures for a Population with Negative FIT Results

Since the data provided from the BCSS does not provide follow up information and diagnoses for those with negative screening test results, exploratory analyses were conducted to provide estimates of test accuracy measures for the FIT used in this sample population. Two approaches were used to calculate estimates of test accuracy. Firstly, the ratio of interval cancers to screen detected cancers for FIT was obtained from a recent

systematic review[55] and applied to the study data to populate a 2 by 2 table. The prevalence of colorectal cancer/polyps from the dataset extracted from GP records for a screening population in **Chapter 5** was also used as a comparison. Lastly, a logit model was used to extrapolate the proportion of cancers/advanced adenomas for those with a screening test of <20 µg/g and this data was used to populate a 2 by 2 table for the study dataset.

3.0 RESULTS

3.1 Study Population

From the total of 40,930 individuals who were sent a FIT kit, 27,066 (66.13%) adequately participated (those who had a definitive positive or negative result) and from this 2,117 (7.82%) had a FIT result of ≥20 µg Hb/g faeces which was classed as positive. From this group, 1818 (85.88%) had a definitive outcome recorded (i.e. not 'cancelled' or 'not attended'). This is a similar proportion of those undergoing further investigation as reported in other studies.[56] Where a diagnostic appointment was made and an individual did not attend, this was classified as 'Not attended' and where an appointment was cancelled the outcome was classified as 'Cancelled' (**Appendix 5**). Eight records were missing IMD and so a final population of 1810 participants were used for multiple logistic regression analysis (**Figure 2** for study flow diagram). The mean age of this group was 66.54 years (See **Table 3** for outcome by age and sex). The FIT result ranged from 20 µg Hb/g faeces to 20,854 µg Hb/g faeces (other studies have reported similarly high results[7]), with a median result of 55.6 µg Hb/g faeces. There were 912 individuals served by the Midlands hub and 898 by the Southern hub.

Seventy-three cancers, 214 high risk adenomas, 262 intermediate risk adenomas and 466 low risk adenomas were detected in the study group. This gave 549 cases with a positive outcome (cancer and advanced adenomas) and 1261 participants with a negative outcome. The FIT concentration increased relative to the severity of the outcome (**Figure 3**) i.e. a normal result had the lowest FIT concentrations on average compared to cancers having the highest FIT results on average.

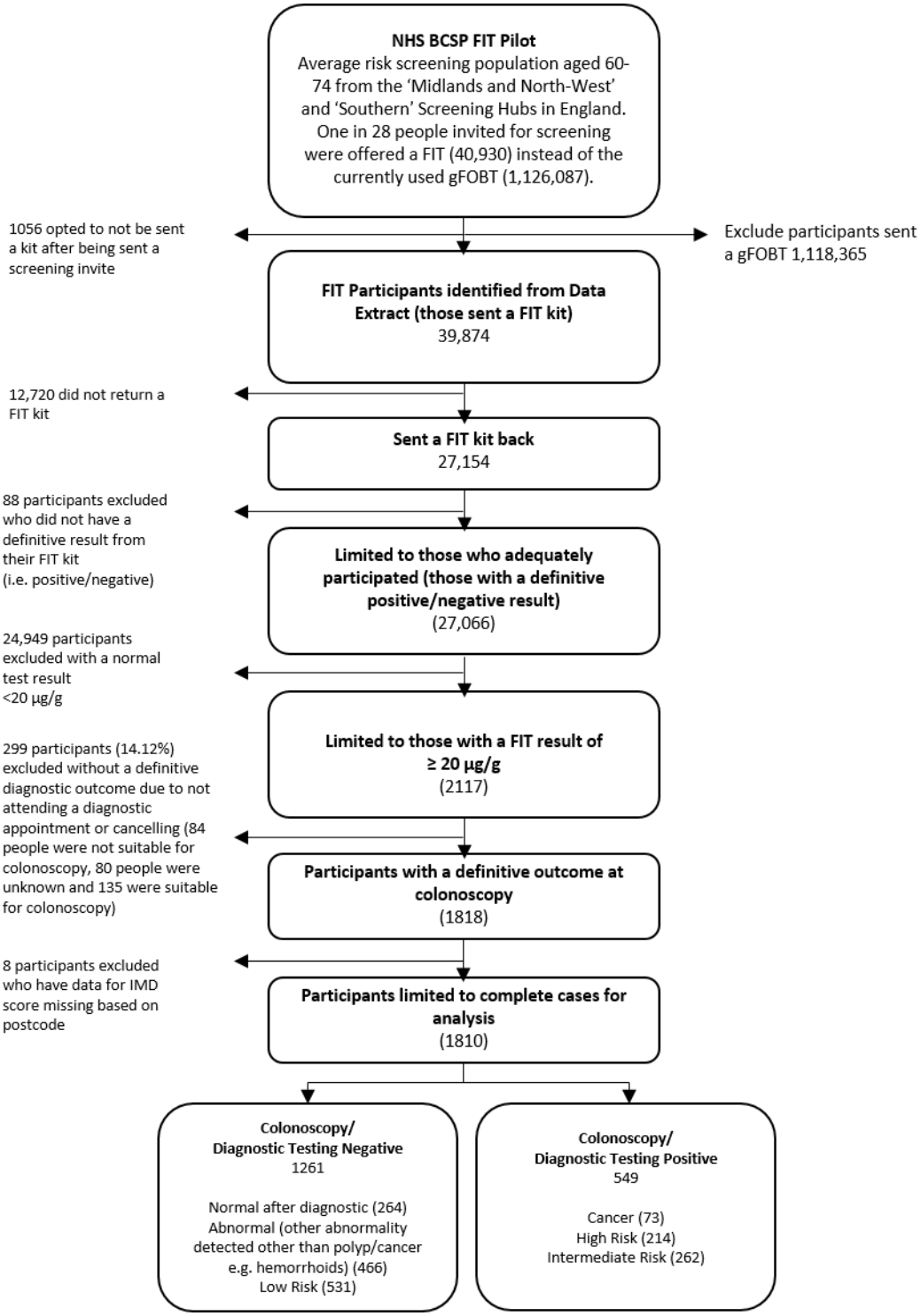


Figure 2: Study flow diagram for the FIT data

Diagnostic outcome by age and sex (n=1810)										
Diagnostic outcome	**Age Category Female**					**Age Category Male**				
	≤60	**61-65**	**66-70**	**71-75**	**Total**	**≤60**	**61-65**	**66-70**	**71-75**	**Total**
Abnormal	29 (10.3%)	90 (31.9%)	93 (33%.0%)	70 (24.8%)	**282 (100%)**	28 (11.2%)	86 (34.5%)	85 (34.1%)	50 (20.1%)	**249 (100%)**
Cancer	2 (8.7%)	5 (21.7%)	6 (26.1%)	10 (43.5%)	**23 (100%)**	4 (8.0%)	15 (30.0%)	21 (42.0%)	10 (20.0%)	**50 (100%)**
High risk Adenoma	7 (10.1%)	12 (17.4%)	27 (39.1%)	23 (33.3%)	**69 (100%)**	17 (11.7%)	37 (25.5%)	58 (40.0%)	33 (22.8%)	**145 (100%)**
Intermediate risk Adenoma	13 (13.1%)	23 (23.2%)	45 (45.5%)	18 (18.2%)	**99 (100%)**	20 (12.3%)	53 (32.5%)	52 (31.9%)	38 (23.3%)	**163 (100%)**
Low risk Adenoma	18 (8.9%)	67 (33.0%)	73 (36.0%)	45 (22.17%)	**203 (100%)**	33 (12.5%)	84 (31.9%)	90 (34.2%)	56 (21.3%)	**263 (100%)**
Normal (No Abnormalities Found)	23 (16.0%)	45 (45.5%)	57 (39.6%)	19 (13.19%)	**144 (100%)**	13 (10.8%)	39 (32.5%)	36 (30.0%)	32 (26.7%)	**120 (100%)**
Total	**92 (11.2%)**	**242 (29.5%)**	**301 (36.7%)**	**185 (22.6%)**	**820 (100%)**	**115 (11.6%)**	**314 (31.7%)**	**342 (34.5%)**	**219 (22.1%)**	**990 (100%)**

Table 3: Diagnostic outcome by age and sex for included participants who adequately participated, had a FIT ≥20 µg Hb/g faeces and with a definitive outcome (n=1810)

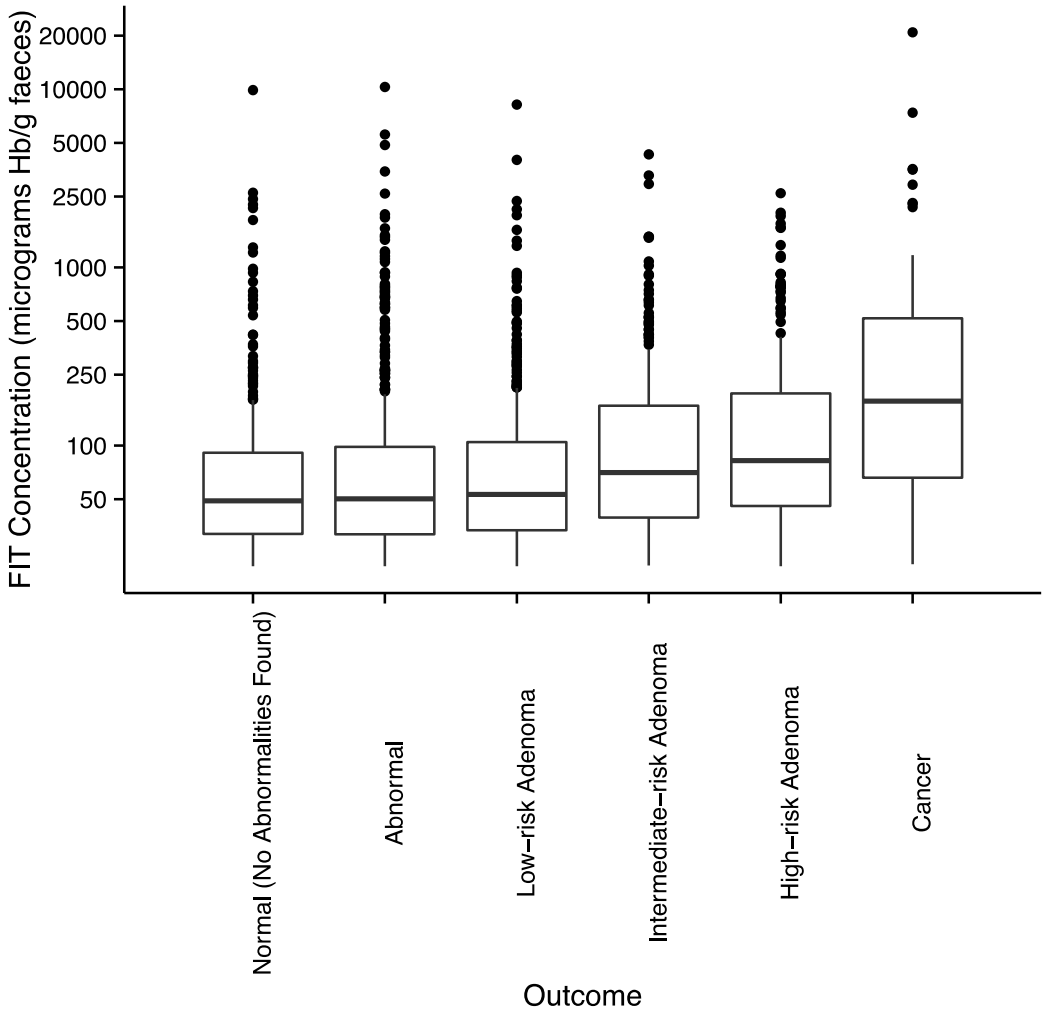


Figure 3: Boxplots of FIT concentration ordered by median for each diagnostic outcome from a normal diagnostic test to the detection of CRC (n=1810). Line is the median, box is the interquartile range, whiskers give 1.5 times the interquartile range and observations outside of this are plotted individually.

3.2 Univariable Logistic Regression (n=2116)

The association of the routine risk indicators with the detection of colorectal cancer and advanced adenomas at colonoscopy are given below using the population who had a FIT result ≥20 µg/g to utilize all available data (n=2116). Risk indicators found to be significant (p value <0.05) in univariable analysis included the FIT result, sex and previous response to screening. The odds of a male being diagnosed at colonoscopy with cancer or advanced adenoma was 1.88 times higher compared to females (OR 1.88, 95% CI: 1.53-2.31). A previous non-responder to screening was 2.7 times more likely of having the outcome diagnosed at colonoscopy compared to a first time screenee at baseline (OR 2.7, 95% CI: 1.77-4.17).

Variable	P-value	Odds Ratio (95% CI)	Missing (n)
Log of FIT result	0.00	1.46 (1.33-1.60)	298 observations deleted due to missingness
Age (at episode start) (continuous)	0.06	1.02 (0.999-1.05)	298
Male (female as base)	0.00	1.88 (1.53-2.31)	298
IMD Score	0.53	0.997 (0.99-1.00)	306
Rank of IMD Score	0.27	1.00 (0.99-1.00)	306
PREVIOUS RESPONSE TO SCREENING			
First Time invitee	-	-	298
Previous non-responder (compared to first time screenee at baseline)	0.00	2.70 (1.77-4.17)	298
Previous responder (compared to first time screenee at baseline)	0.03	1.50 (1.05-2.19)	298
Southern hub (compared to Midlands at baseline)	0.64	1.05 (0.86-1.28)	298

Table 4: Univariable logistic regression of BCSS routine data with colorectal cancer/advanced adenoma detected at colonoscopy.

3.3 Multivariable Logistic Regression (n=1810)

The logistic regression model with FIT only is given in **Table 5.** Complete cases were used for multiple logistic regression analysis (n=1810). Backwards elimination with cross validation was used to build a multivariable logistic regression model (**Table 6**)**.** The p-values for the likelihood ratio test were analysed to determine whether a variable should be dropped from the model using cross-validation. Risk indicators found to be significant and retained in the model with a p value of <0.1 were the FIT result, sex and previous screening history. The full risk equation is given below in **Equation 1**.

The odds of colorectal cancer and advanced adenoma increase as the FIT result increases (OR: 1.434, CI: 1.309 – 1.573), for males (OR: 1.749, CI: 1.415 – 2.166) and for previous non-

responders (OR: 2.271, CI: 1.422 – 3.667). Age was not found to have a statistically significant influence on whether cancer was detected, but was retained in the model due to clinical importance and it was found to be a significant predictor in the literature and **Chapter 2** (OR: 1.020, CI: 0.889 – 2.112).[7 8 43] All possible pairwise interactions were investigated due to the small number of predictors retained, none of which were significant at the 5% level.

Statistical power calculations for reliable predictions frequently consider the rule of thumb of 10 events per variable (EPV) as demonstrated in two simulation studies.[19 57 58] The events per variable/parameter for this model were 91.5 (549 events/5 parameters plus 1 for the parameter representing the constant). This result satisfies the rule and suggests that there are enough EPVs to accurately estimate the coefficients in the logistic regression model. Prediction models with EPVs below ten can be overfitted to the data.

Coefficients	Estimate	Std. Error	Bootstrapped coefficient bias	Bootstrapped standard error	Pr (>|z|)	OR 95% CI
Intercept	-2.487	0.212	-0.005	0.217	<0.001	0.083 (0.055 – 0.126)
log(FIT Result +1)	0.374	0.046	0.001	0.048	<0.001	1.454 (1.329 – 1.592)

Null deviance – 2221.4 on 1809 degrees of freedom **Residual deviance** – 2153.6 on 1808 degrees of freedom **AIC** – 2157.6
Number of Fisher Scoring Iterations - 4

Table 5: FIT only Logistic Regression Model

Coefficients	Estimate	Std. Error	Bootstrapped coefficient bias	Bootstrapped std. error	Pr (>|z|)	OR 95% CI
Intercept	-4.439	0.934	-2.167e-02	0.949	<0.001	0.012 (0.002 – 0.073)
log(FIT Result +1)	0.360	0.047	1.963e-03	0.049	<0.001	1.434 (1.309 – 1.573)
Age at episode start	0.020	0.015	6.691e-05	0.015	0.171	1.020 (0.991-1.050)
Sex (male)	0.559	0.109	1.360e-03	0.108	<0.001	1.749 (1.415-2.166)
First Time Invitee	0.000	-	-	-	-	-
Previous non responder (compared to first time screen)	0.820	0.241	3.818e-03	0.245	0.001	2.271 (1.422-3.667)
Previous responder (compared to first time screen)	0.308	0.220	6.516e-03	0.220	0.162	1.361 (0.889 – 2.112)

Null deviance – 2221.4 on 1809 degrees of freedom **Residual deviance** – 2103.0 on 1804 degrees of freedom
AIC – 2115 Number of Fisher Scoring Iterations – 4 **Nagelkerke's R^2** - 0.09 (risk-adjusted model) and 0.05 (FIT only)
Events per variable/parameter - 91.5 (549 events/5 parameters plus 1 for the parameter representing the constant)

Table 6: Multiple Logistic Regression Model (FIT combined with risk indicators)

Risk Equation:

$$\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = a + \beta X_1 + \beta X_2 + \cdots$$

$$p = \frac{e^{a+\beta X_1+\beta X_2+\cdots}}{1 + e^{a+\beta X_1+\beta X_2+\cdots}}$$

p =probability
a =constant
β =coefficient predictor variable

$$\boldsymbol{p = \frac{e^{-4.44+0.360X_1+0.02X_2+0.56X_3+0.82X_4+0.31X_5}}{1 + e^{-4.44+0.360X_1+0.02X_2+0.56X_3+0.82X_4+0.31X_5}}}$$

p =probability
a =constant
β =coefficient predictor variable
x_1 = log(FIT Result +1)
x_2 =Age at episode start
x_3 =Sex (male)
x_4 = Previous non responder (compared to first time screen)
x_5 = Previous responder (compared to first time screen)

Equation 1: Risk equation for the multivariable model combining FIT with other risk predictors.

3.4 Overall Model FIT

The deviance for the FIT only model was 2153.6 on 1808 degrees of freedom (cross validated deviance 2157.8), the risk adjusted model improved this measure giving a value of 2103.0 on 1804 degrees of freedom (cross validated deviance 2113.995). The likelihood ratio test was used to compare the goodness of fit of the nested models by assessing the difference in deviance. The risk-adjusted model had a significantly better fit compared to FIT only: $\chi^2(4) = 50.57$, $p<0.001$).

The overall model fit for the logistic regression models was also assessed using pseudo R^2 measures (**Table 7**). Cox and Snell's R^2 assesses the improvement of the full model over the intercept model and gives a value of 0.037 for the FIT only model compared to 0.063 for the risk-adjusted model. Nagelkerke's R^2 was 0.090 for the risk adjusted model compared with 0.052 for the FIT only model.

Pseudo R^2	FIT only Model	Risk adjusted Model
McFadden	0.031	0.053
Adj.McFadden	0.028	0.047
Cox.Snell	0.037	0.063
Nagelkerke	0.052	0.090
McKelvey.Zavoina	0.049	0.090
Effron	0.038	0.065
Count	0.696	0.701
Adj.Count	-0.004	0.015
AIC	2157.6	2115.0
Corrected.AIC	2157.6	2115.1

Table 7: Pseudo R^2 measures for the multivariable logistic regression model (FIT combined with risk indicators)

3.5 Calibration

Calibration assesses the agreement between the observed and expected risk. The calibration plots of observed risk against predicted risk are given for both models in **Figure 4** where the study group is split into deciles according to their predicted probabilities. The Hosmer-Lemeshow goodness of fit test is used to assess calibration. The calibration for the risk-adjusted model was 0.898 versus 0.481 for the FIT (**Table 8** and **Table 9**). Small p-values and points which are far from the line of equality in the calibration plot indicate a poor fit. A well calibrated model has predictions around the 45 degree line on the calibration plot with a slope close to 1 and intercept close to 0. Since the number of group splits for the predicted probabilities can affect the corresponding Hosmer and Lemeshow p values, group divisions between 5-15 and their p-values are presented in **Appendix 6**. The slope/gradient of the FIT only model based on the calibration plot was 1.005x and for the risk-adjusted model 1.028x. The closer the calibration slope to 1, the better the fit of the model.

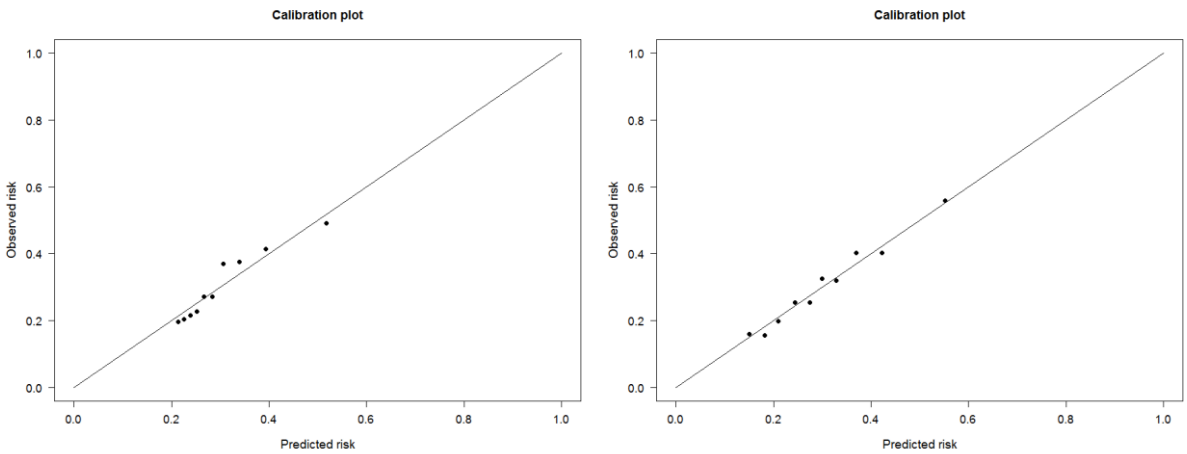


Figure 4: Calibration plot of observed risk versus predicted risk for FIT only (left) and risk-adjusted FIT models (right).

Model 2 risk adjusted Calibration	Total	Mean of the predicted probabilities	Mean of the observed probabilities	Predicted number of cases	Observed number of cases
[0.105,0.171)	181	0.150	0.160	27.18	29
[0.171,0.194)	181	0.182	0.155	32.97	28
[0.194,0.227)	181	0.210	0.199	37.97	36
[0.227,0.260)	181	0.244	0.254	44.24	46
[0.260,0.285)	181	0.274	0.254	49.55	46
[0.285,0.313)	181	0.299	0.326	54.17	59
[0.313,0.346)	181	0.329	0.320	59.51	58
[0.346,0.392)	181	0.369	0.403	66.84	73
[0.392,0.462)	181	0.423	0.403	76.59	73
[0.462,0.843]	181	0.552	0.558	99.98	101

Chi squared – 3.52 **degrees of freedom** – 8 **p value** - 0.8977

Table 8: Observed versus expected risk for the Hosmer-Lemeshow goodness of fit test using deciles of risk for the risk-adjusted model

Model 1 FIT only Calibration	Total	Mean of the predicted probabilities	Mean of the observed probabilities	Predicted number of cases	Observed number of cases
[0.206,0.219)	184	0.213	0.196	39.15	36
[0.219,0.233)	182	0.225	0.203	41.04	37
[0.233,0.245)	177	0.239	0.215	42.30	38
[0.245,0.259)	185	0.252	0.227	46.60	42
[0.259,0.274)	181	0.267	0.271	48.37	49
[0.274,0.294)	177	0.283	0.271	50.12	48
[0.294,0.320)	181	0.306	0.370	55.38	67
[0.320,0.363)	181	0.339	0.376	61.40	68
[0.363,0.434)	181	0.393	0.414	71.15	75
[0.434,0.775]	181	0.517	0.492	93.49	89

Chi squared – 7.53 **degrees of freedom** – 8 **p value** - 0.4807

Table 9: Observed versus expected risk for the Hosmer-Lemeshow goodness of fit test using deciles of risk for the FIT only model

3.6 Discrimination

Discrimination was measured by assessing the area under the ROC curve. This measure determines how well the model can discriminate between those with and those without the outcome of interest. The ROC curves for both models are presented below. The AUC for the FIT only model was 0.63 (95% CI: 0.60 - 0.66) compared to 0.66 for the risk-adjusted model indicating improved discrimination (95% CI: 0.63 - 0.69). A ROC test using 10,000 bootstrap iterations shows that the AUCs are significantly different (D = -2.7601, p-value = 0.006).

3.7 Predictiveness Curve

The predictiveness curve for the risk-adjusted model is presented in **Figure 5**. The predictiveness curve integrates two statistical approaches to model evaluation; modelling the risk of disease and classification performance. This is achieved by displaying information on both classification and predictiveness within one plot. A risk threshold of 0.389 using the risk-adjusted model is equivalent to a FIT cut-off of 160 µg/g (setting the recall rate the same). With a risk probability of 0.389, around 21% of patients in the cohort have risk probabilities at or above this threshold using the risk-adjusted model. For FIT only on the other hand around 18% have risk probabilities above this threshold. The plot summarizes the distribution and range of risk probabilities from the model and from FIT only. A reference line can be included at 0.30, which represents the prevalence of disease in the sample, to correspond to a predictiveness curve for an uninformative risk model (assigning subjects at the same risk).[48]

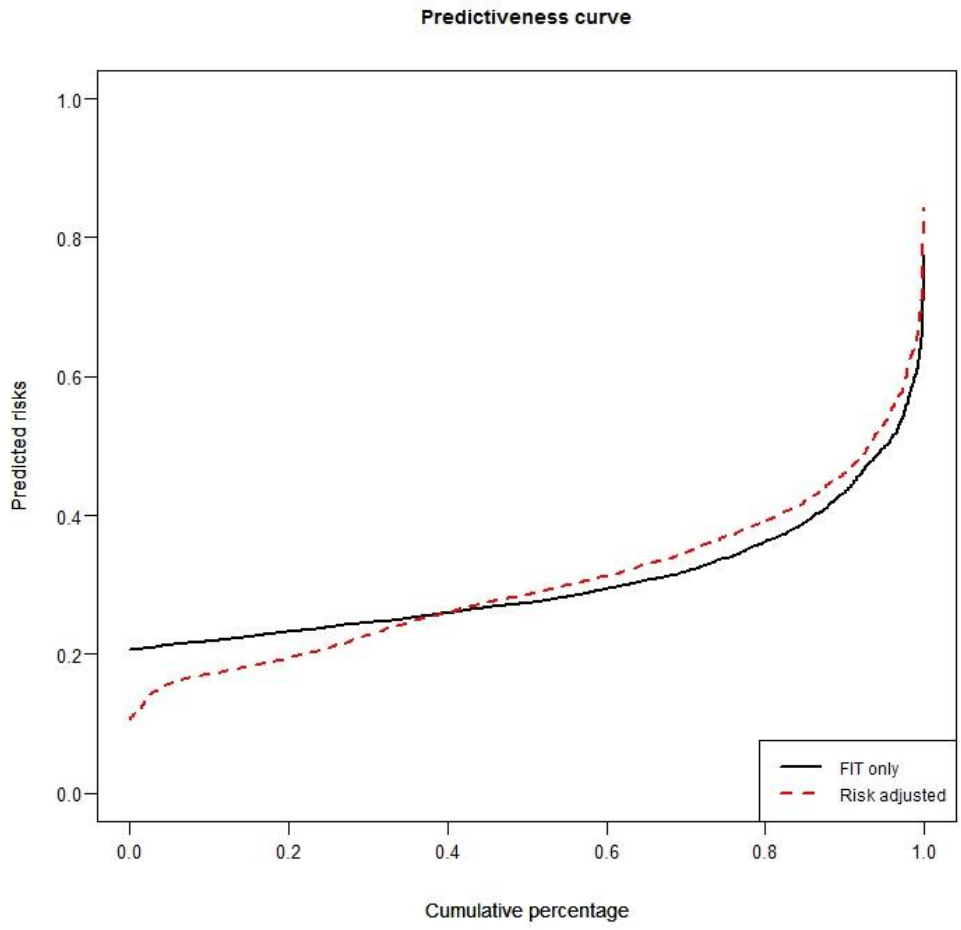


Figure 5: Predictiveness curves for the FIT only model and Risk adjusted model

3.8 Test Accuracy

Test accuracy is presented in a 2 by 2 table for a threshold of 160 µg/g in **Table 10**. At all investigated thresholds (30-180 µg Hb/g faeces), the sensitivity and specificity of risk adjusted FIT is greater than FIT alone (see **Table 11**). Individuals were sorted by predicted probability and the number of referrals kept the same between using the FIT alone and using risk adjusted FIT. For instance, for a FIT cut-off of 160 µg/g, 375 individuals had a positive result and for an equivalent threshold of risk (predicted probability of 0.389) 375 individuals would also be referred when applying the logistic regression model. At this threshold, the FIT has a sensitivity of 30.78% versus 33.15% for the risk-adjusted model and a specificity of 83.66% versus 84.69%. The ROC curves are given for both models in **Figure 6** which displays the sensitivity and specificity pairs for all available thresholds.

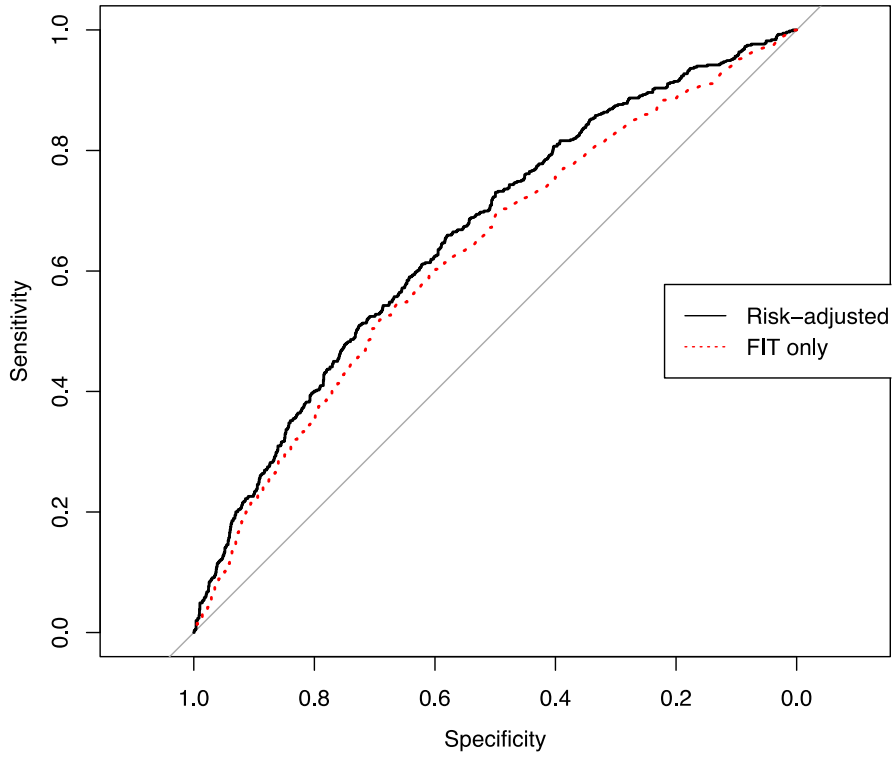
The risk-adjusted model for this sample population leads to the detection of 13 additional advanced adenomas and the same number of cancers (17 more high risk adenomas, 4 fewer intermediate risk adenomas) when compared to the FIT only at an equivalent threshold of 160 µg/g. The severity profiles of the detected lesions are shown in **Table 10** (further thresholds are presented in **Chapter 4**). The risk-adjusted model therefore improves the diagnostic yield of high-risk adenomas. At the same time, there is a reduction in the number of false positives and negatives as well as an increase in the number of true negatives.

2 by 2 table for FIT only and the risk-adjusted logistic regression model					
160 µg Hb/g faeces Threshold	Diagnostic Positive		Diagnostic Negative		Total
	FIT	Risk-adjusted	FIT	Risk-adjusted	
FIT/Risk Positive	**169** *37 - Cancer* *66 - High Risk Adenoma* *66 - Intermediate Risk Adenoma*	**182** *37 - Cancer* *83- High Risk Adenoma* *62 - Intermediate Risk Adenoma*	**206** *70 - Abnormal* *92 - Low Risk Adenoma* *44 - Normal (No Abnormalities Found)*	**193** *69 - Abnormal* *81 - Low Risk Adenoma* *43 - Normal (No Abnormalities Found)*	**375**
FIT/Risk Negative	**380** *36 - Cancer* *148 - High Risk Adenoma* *196 - Intermediate Risk Adenoma*	**367** *36 - Cancer* *131 - High Risk Adenoma* *200 - Intermediate Risk Adenoma*	**1055** *396 - Abnormal* *439 - Low Risk Adenoma* *220 - Normal (No Abnormalities Found)*	**1068** *397 - Abnormal* *450 - Low Risk Adenoma* *221 - Normal (No Abnormalities Found)*	**1435**
Total	**549**		**1261**		**1810**
FIT only: Sensitivity 30.78%, Specificity 83.66%, PPV 45.07%, NPV 73.52%, FIT positivity 20.72%, Cancer Detection Rate 9.34% **Risk adjusted: Sensitivity 33.15%, Specificity 84.69%, PPV 48.53%, NPV 74.42%, FIT positivity 20.72%, Cancer Detection Rate 10.60%**					

Table 10: 2 by 2 table for FIT only and the risk-adjusted logistic regression model. A threshold of 160 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.389 for the risk-adjusted model. Profiles of outcome severity are also given.

Clinical sensitivity and specificity pairs for FIT thresholds between 30 and 180 µg Hb/g faeces and the corresponding risk thresholds.			
Model	FIT (µg Hb/g faeces)/ Risk Threshold (probability)	Sensitivity (%)	Specificity (%)
FIT only **Risk-adjusted LR**	30.00 0.191	88.34 90.35	22.20 23.08
FIT only **Risk-adjusted LR**	40.00 0.242	76.68 80.15	38.94 40.44
FIT only **Risk-adjusted LR**	50.00 0.272	69.03 70.86	50.04 50.83
FIT only **Risk-adjusted LR**	60.00 0.295	60.66 62.48	59.24 60.03
FIT only **Risk-adjusted LR**	70.00 0.310	55.19 57.19	64.63 65.42
FIT only **Risk-adjusted LR**	80.00 0.321	51.18 52.64	69.31 69.94
FIT only **Risk-adjusted LR**	90.00 0.336	45.72 48.63	72.56 73.83
FIT only **Risk-adjusted LR**	100.00 0.346	42.44 44.99	75.26 76.37
FIT only **Risk-adjusted LR**	110.00 0.356	40.07 42.99	77.08 78.35
FIT only **Risk-adjusted LR**	120.00 0.362	38.07 40.26	78.59 79.54
FIT only **Risk-adjusted LR**	130.00 0.371	34.79 37.89	80.33 81.68
FIT only **Risk-adjusted LR**	140.00 0.379	33.70 36.25	81.60 82.71
FIT only **Risk-adjusted LR**	150.00 0.383	32.42 35.15	82.39 83.58
FIT only **Risk-adjusted LR**	160.00 0.389	30.78 33.15	83.66 84.69
FIT only **Risk-adjusted LR**	170.00 0.392	29.87 31.69	84.30 85.09
FIT only **Risk-adjusted LRs**	180.00 0.399	28.60 30.05	85.57 86.20

Table 11: Clinical sensitivity and specificity pairs for FIT thresholds between 30 and 180 µg Hb/g faeces and the corresponding risk thresholds for the logistic regression model.



AUC (95% CI) for the Risk-adjusted Logistic Regression Model: 0.659 (0.632 - 0.686)
AUC (95% CI) for the FIT only: 0.628 (0.600 - 0.656)

Figure 6: ROC curves for FIT only compared to risk-adjusted FIT

3.9 Test Accuracy for Subgroups

Presenting the results by sex, **Table 12** shows the risk model at 160 µg/g recalls more men and fewer women, increases detection in men but decreases detection in women in comparison to the FIT result alone. The FIT result alone recalled 225 men (115 TP – true positives, 110 FP – false positives) of which 115 had cancer or advanced adenoma (51.11%), and 150 women (54 TP, 96 FP) where 54 (36%) had cancer or advanced adenoma. The logistic regression model recalled 314 men (156 TP, 158 FP) of which 156 (49.68%) had cancer or advanced adenoma, and 61 women (26 TP, 35 FP) where 26 (42.62%) had cancer or advanced adenoma. Detection rates by screening history and sex subgroup are shown in **Table 13**. There is a reduction in the detection rate for female first time invitees and the detection rate for male previous non-responders more than doubled (16.98% to 37.11%).

2 by 2 table for FIT only and the risk-adjusted logistic regression model split by sex.										
160 µg Hb/g faeces Threshold	Diagnostic Positive	FIT only		Risk-adjusted		Diagnostic Negative	FIT only		Risk-adjusted		Total
		Male	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	
FIT/Risk Positive	***Total***	**115**	**54**	**156**	**26**	***Total***	**110**	**96**	**158**	**35**	**375**
	Cancer	27	10	29	8	***Low Risk Adenoma***	41	29	60	9	
	High risk Adenoma	45	21	72	11	***Abnormal***	51	41	66	15	
	Intermediate risk Adenoma	43	23	55	7	***Normal (No Abnormalities Found)***	18	26	32	11	
		Male	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	
FIT/Risk Negative	***Total***	**243**	**137**	**202**	**165**	***Total***	**522**	**533**	**474**	**594**	**1435**
	Cancer	23	13	21	15	***Low Risk Adenoma***	222	174	203	194	
	High risk Adenoma	100	48	73	58	***Abnormal***	198	241	183	267	
	Intermediate risk Adenoma	120	76	108	92	***Normal (No Abnormalities Found)***	102	118	88	133	
Total		**549**					**1261**				**1810**

Table 12: 2 by 2 table for FIT only and the risk-adjusted logistic regression model split by sex. A threshold of 160 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.389 for the risk-adjusted model. Profiles of outcome severity are also given.

Cancer/advanced adenoma detection rate by screening history and sex subgroup										
	FIT Only					Risk Model				
Subgroup	TP	FP	FN	TN	Cancer/Advanced Adenoma Detection Rate (%)	TP	FP	FN	TN	Cancer/Advanced Adenoma Detection Rate (%)
Female First Time Invitee	4	10	12	64	4.44	0	1	16	73	0.00
Male First Time Invitee	12	14	13	61	12.00	5	6	20	69	5.00
Female Non Responder	14	10	18	49	15.38	14	12	18	47	15.38
Male Non Responder	27	15	47	70	16.98	59	63	15	22	37.11
Female Responder	36	76	107	420	5.63	12	22	131	474	1.88
Male Responder	76	81	183	391	10.40	92	89	167	383	12.59
TP – True Positive; FP – False Positive; FN- False Negative; TN – True Negative										

Table 13: Cancer/advanced adenoma detection rate by screening history and sex subgroup (Threshold 160 µg Hb/g) for the FIT only and risk-adjusted model.

3.10 Additional Predictors and their effect on FIT positivity

For the sample population that had adequately participated with a definitive FIT result of positive or negative (n=27,066) there were 498 missing results for sample return time and mean maximum temperature. This is most likely due to the participant not filling out the sample date on the FIT kit properly. This was a relatively low number of individuals so these were excluded from the unvariable analysis. 26,614 participants had a negative result and 452 had a positive result and there were 14,305 females (52.85%) and 12,761 males. The mean return time was 2.14 days (SD: 1.26) and the mean maximum ambient temperature was 18.92°C (SD: 3.64) (boxplots for these variables are provided in **Appendix 7**). Time series plots for mean maximum temperatures are shown for both the Midlands and Southern hubs in **Figure 7** and **Figure 8**.

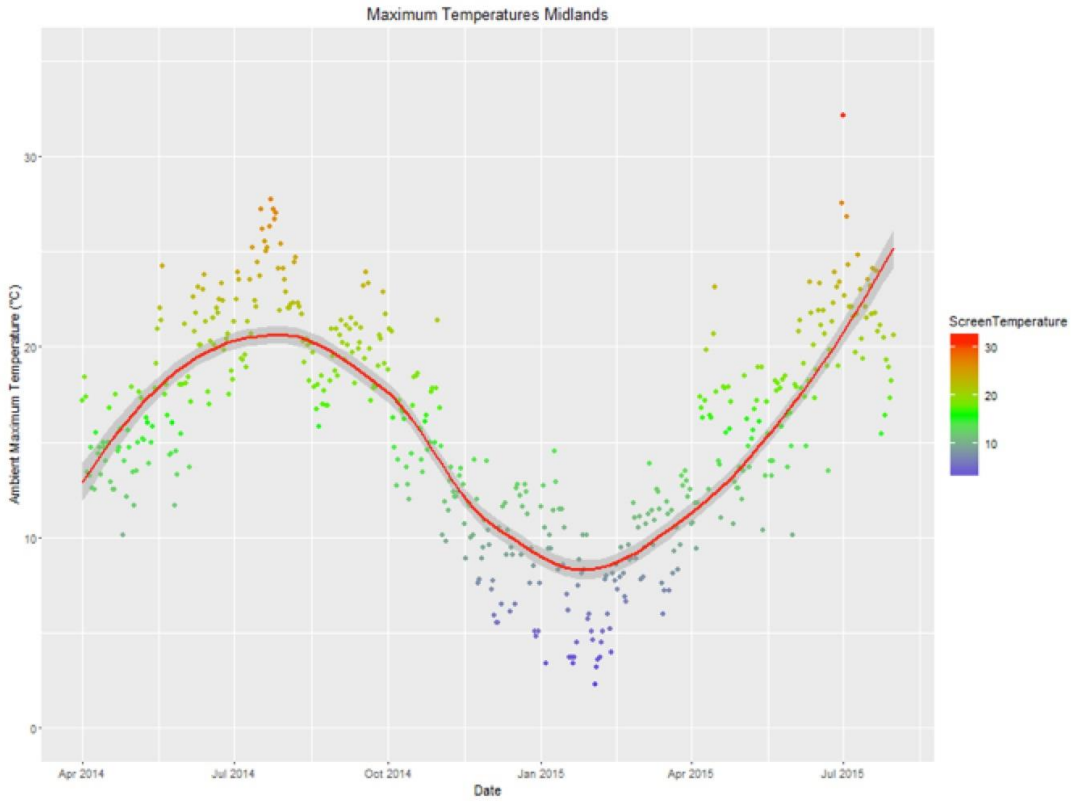


Figure 7: Time series plot showing the mean maximum temperature recorded each day for the Midlands Hub from April 2014 to July 2015.

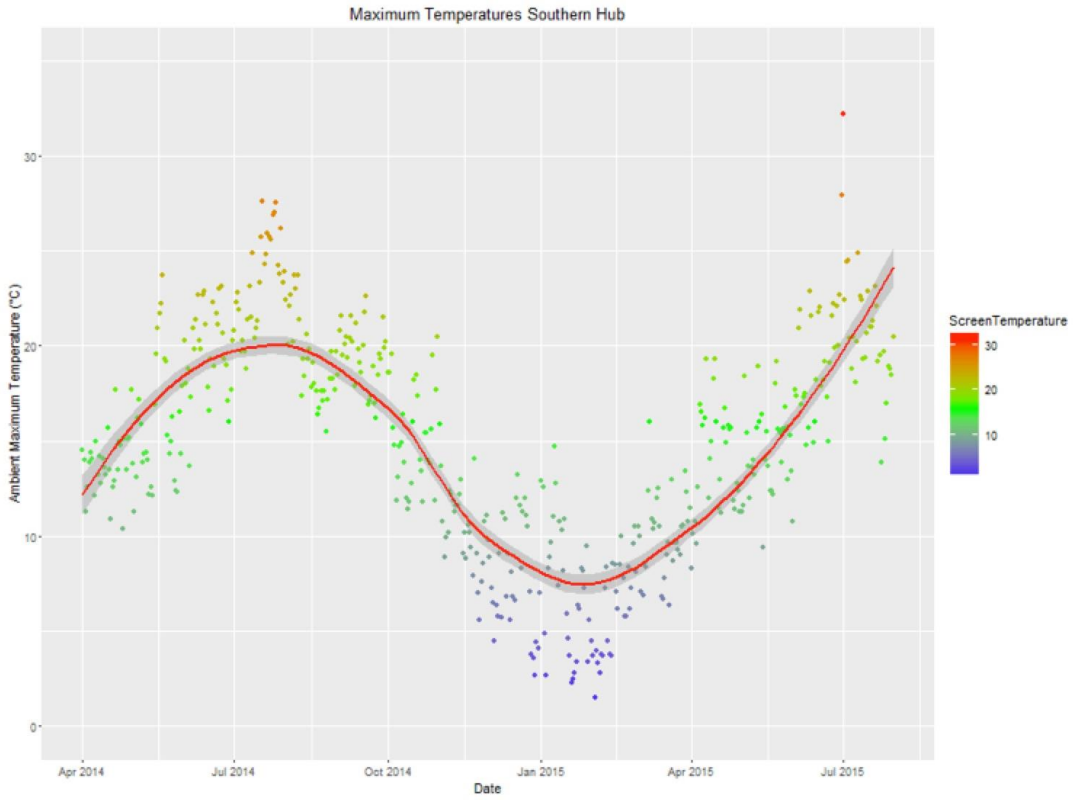


Figure 8: Time series plot showing the mean maximum temperature recorded each day for the Southern Hub from April 2014 to July 2015.

Univariable logistic regression at a threshold of 160 µg Hb/g faeces (**Table 14**) shows that sample return time is statistically significant and associated with a positive FIT result. As sample return time increases by 1 day, the odds of FIT positivity increases by 0.105. This result was still statistically significant when adjusting for sex, previous screening history and age (p=0.003). Sex was significant, even after adjusting for screening history and age (p<0.001). IMD was still significant after adjusting for sex, previous screening history and age (p<0.001). Conversely, the ambient temperature that FIT is exposed to was not statistically significant. The screening hub becomes no longer significant when adjusting for IMD, Sex and age (p=0.326) as these factors would explain underlying differences in the populations served by the hubs.

Variable	P-value	Odds Ratio (95% CI)	Missing (n)
Sample Return Time	<0.001	1.10 (1.04-1.17)	498 observations deleted due to missingness
Ambient Mean Maximum Temperature	0.467	0.99 (0.96-1.02)	498
Sex (males compared to females at baseline)	<0.001	1.84 (1.52-2.24)	0
Age	0.016	1.03 (1.01-1.05)	0
First Time invitee	-	-	298
Previous non-responder (compared to first time screenee at baseline)	<0.001	2.55 (1.79-3.68)	0
Previous responder (compared to first time screenee at baseline)	0.09	1.31 (0.97-1.80)	0
IMD Score	<0.001	1.01 (1.01-1.02)	200
Screening Hub (southern hub compared to midlands at baseline)	0.037	0.82 (0.68-0.99)	0

Table 14: Univariable logistic regression results for a threshold of 160 µg/g.

3.11 Estimation of Test Accuracy Measures for a Population with a Negative FIT Result

This section describes exploratory analyses to provide estimates of test accuracy measures for the FIT used in this sample population (in particular the NPV). Ideally, to determine unbiased diagnostic accuracy measures, all individuals who are screened should be tested with a diagnostic reference standard. This is a common limitation for population based screening studies which have large sample populations but do not offer diagnostic testing to individuals below a screening test cutoff (in this case 20 µg Hb/g faeces) due to cost effectiveness. This leads to partial verification bias where there is selectively missing data on disease outcome causing biased estimates of test accuracy.[59] A sample population with a screening test of 20 µg Hb/g faeces and above will result in a population at higher risk of having colorectal cancer detected, potentially inflating test accuracy measures such as the sensitivity.

Alternatively, two year follow up data can be used for those with a negative screening test result, linking screening data to the cancer registry. This too comes with its limitations due to the different reference standards used for diagnosis (colonoscopy versus two year follow up) leading to differential verification bias.[59] Screening for instance is more likely to identify slow growing disease compared to those which present symptomatically (length time bias).[60]

An additional consideration when investigating false negative results or interval cancers are the different definitions of interval cancers which may be used in each study. Screening programmes often define interval cancers as those missed at colonoscopy and by the screening test.[61] Interval cancers can be considered different to a false negative screening result which is defined at the time of colonoscopy (a shorter term outcome).

Two approaches were considered to estimate test accuracy for the sample population; (i) Modelling the available data to extrapolate the proportion of cancers for those with a screening test of less than 20 µg Hb/g faeces and (ii) Examining data available in the literature.

Examining the Literature

A recent systematic review by Wieten *et al*.[55] identified that the incidence rate for colorectal interval cancers is lower when using the FIT (20 interval cancers per 100,000 person years) compared to the guaiac based test (34 interval cancers per 100,000 person years). The findings suggest that for each interval cancer missed by the FIT, 2.6 screen detected colorectal cancers are diagnosed, whereas for the gFOBT the ratio is 1 to 1.2. Based on these figures, the following 2 by 2 table can be estimated using the data from this study, whereby screen detected cancers have a ratio of 2.6 to 1 interval cancer (**Table 15**).

The ratio may be underestimated because it did not include advanced adenomas in addition to cancers but provides a reasonable estimate for these data. The sensitivity found from this estimate was 0.72 along with a specificity of 0.95 which matches reasonably closely the pooled estimate 0.79 (95% CI, 0.69 to 0.86) from the systematic review of faecal immunochemical test accuracy by Lee *et al.*[62] The PPV was 0.30 which is specific to the prevalence of the sample population and the NPV was estimated as 0.99. The prevalence of colorectal cancer and advanced adenomas in the study was 2.84% based on these estimated values.

	Positive Diagnostic Result	**Negative Diagnostic Result**	**Total**
FIT positive ≥ 20 µg/g	549	1261	1810
FIT negative < 20 µg/g	212	24,737	24949
Total	761	25,998	26,759
Sensitivity: 0.72%, Specificity: 0.95%, PPV: 0.30, NPV: 0.99, Sample prevalence: 2.84%, FIT positivity 6.76%. **Excluded**: Those with positive test result (≥20 µg/g) but with no diagnostic outcome (27,066 – 307 = 26,759). This will most likely underestimate the number of true positive results.			

Table 15: 2 by 2 table using estimated interval cancer ratios from the literature,[55] and applying to the data obtained from this study.

In a randomised population based screening study in the Netherlands, where everyone obtained a colonoscopy regardless of FIT result, the sensitivity was 31% (95% CI: 23-40) and specificity 97% (95% CI: 96-98) for advanced neoplasia. The PPV for advanced adenoma was 52% (95% CI: 40-64) and NPV 93% (95% CI: 91-94), both these measures are dependent on the prevalence in the sample population. The FIT positivity was 6% using a cut-off of 20 µg/g and sample prevalence for advanced neoplasia was 9% and for CRC 0.6%. The prevalence appears higher since the population tested in this study were screening naïve, whereas the FIT pilot data includes a mix of individuals who have had a different number of screening rounds. As screening rounds continue, the prevalence in advanced neoplasia will drop over time. As a comparator, the Office for National Statistics data for

adults aged 60-74 in England in 2016 gave 0.997% for newly diagnosed cases of colorectal cancer.[63] A study in Scotland looking at interval cancers 2 years after the last negative FIT result (80µg/g used as the threshold) had 31 interval cancers out of 30,893 participants with a definitive test result and 30 screen detected cancers giving a sample prevalence of 0.197%.[64]

Using the data obtained for a screening cohort extracted from GP records investigated in **Chapter 5** can provide an additional comparison. **Table 16** below shows a 2 by 2 table of colorectal cancer or polyp diagnosis by guaiac faecal occult blood test result for participants with 2 years of follow up. This data provides an overall estimate for the prevalence of cancer/polyps in an average risk screening population which is 3.55%. This estimate is higher than the previous estimated sample population prevalence (2.84%) and is presumably due to the inclusion of colorectal polyps as a diagnosis which can reflect less severe disease. The sensitivity and specificity are specific to the guaiac based test and match systematic review estimates reasonably closely (0.47 sensitivity, 0.92 specificity).[65]

gFOBT result	Cancer /Polyp Diagnosis Positive	Cancer /Polyp Diagnosis Negative	Total
Positive	220 colorectal cancers 329 polyps =549	1,084	1633
Negative	226 colorectal cancers 297 polyps =523	28,031	28,554
Total	1072	29,115	30,187
Sensitivity: 0.51, Specificity: 0.96, PPV: 0.34, NPV: 0.98, Sample prevalence: 3.55%, gFOBT positivity 5.41%. N=30,187			

*Table 16: 2 by 2 table of colorectal cancer/polyp diagnosis by guaiac faecal occult blood test result for participants with 2 years of follow up (data from the study reported in **Chapter 5**).*

By applying the prevalence estimate of cancer/polyps to the FIT data in the current chapter (3.55%-2.05% = 1.50% remaining) the following results are obtained; Sensitivity 0.58, specificity 0.95, PPV 0.30, NPV 0.95. The NPV could be potentially underestimated due to the inclusion of colorectal polyps as a diagnostic outcome. This approach may also underestimate the sensitivity of the faecal immunochemical test (0.58) as the ratio of screen detected to interval cancers is based on gFOBT data.

	Positive Diagnostic Result	Negative Diagnostic Result	Total
FIT positive ≥ 20 µg/g	549	1,261	1810
FIT negative < 20 µg/g	402	24,547	24,949
Total	951	25,808	26,759
Sensitivity: 0.58, Specificity: 0.95, PPV: 0.30, NPV: 0.98, Sample prevalence: 3.55%, FIT positivity 6.76%. **Excluded**: Those with positive test result (≥20 µg/g) but with no diagnostic outcome (27,066 – 307 = 26,759). This will likely underestimate the number of true positive results.			

Table 17: 2 by 2 table estimating the number of false negative and true negative results using the sample prevalence obtained from Table 16.

Modelling the available data to estimate the proportion of cancers for those with a result of less than 20 µg/g

A model predicting cancer/advanced adenoma for those with a result of 20 µg/g or over could be used to extrapolate the expected proportion of those with undiagnosed cancer/advanced adenoma for a result of under 20 µg/g. Several assumptions would need to be made to produce such an estimate and different model types (e.g. logit, probit etc) could be investigated to determine which model has the best fit to the nature of the data.

To illustrate such an approach, a logit model was fitted to the FIT result to predict a cancer/advanced adenoma outcome for those with a result of 20 µg/g or over. This sample population which has both a FIT result and outcome (n=1810) was split into deciles based on the FIT result and the logit model was fit over the top and illustrated in **Figure 9**. From this plot, the proportion of those with potential undiagnosed cancer/advanced adenomas for varying FIT results can be estimated. Alternatively the underlying model equation can be used to calculate more accurate figures.

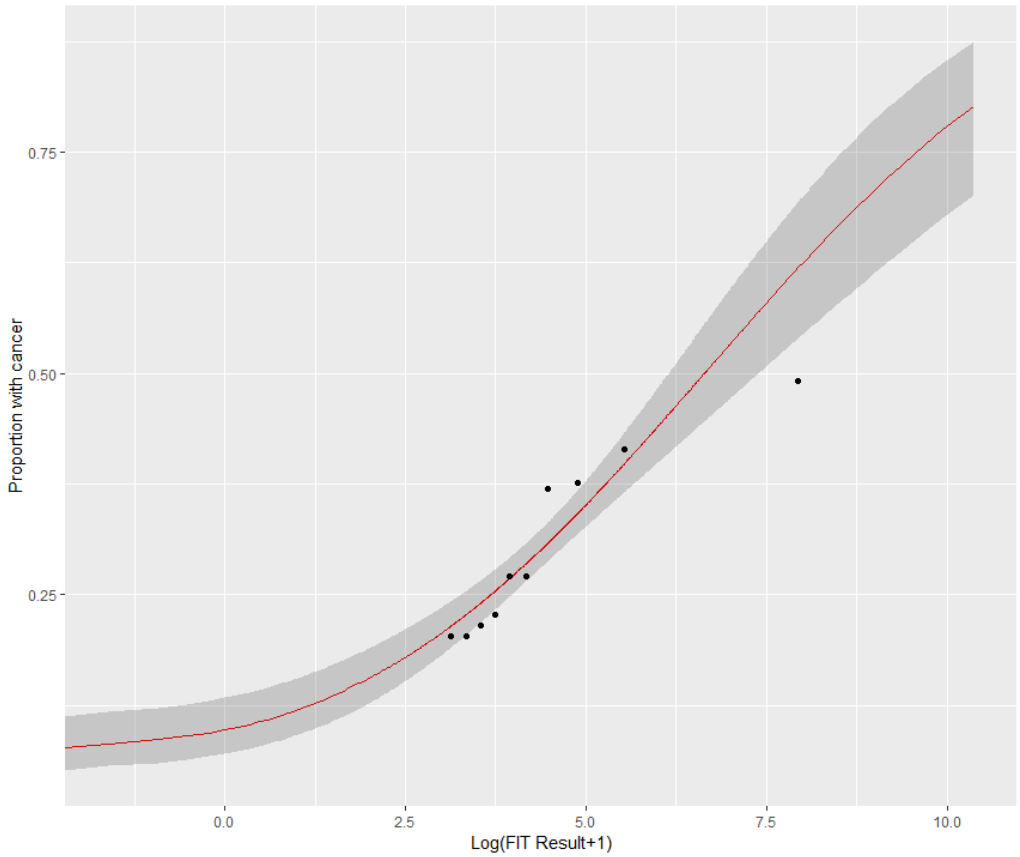


Figure 9: Log of the FIT result plus 1 plotted against the estimated proportion of those with cancer based on the logit model developed using participant data with a FIT result of 20 µg/g and over with a diagnostic outcome and extrapolated for FIT results less than 20 µg/g (cancer here includes both advanced adenomas and colorectal cancers).

To estimate the proportion of those with a FIT result of less than 20 µg/g who might have undiagnosed cancer/advanced adenoma, the proportion of people with a certain FIT result (between 1 to 19) out of those with a result of less than 20 µg/g (n=24,930) was multiplied by the probability of cancer/advanced adenoma based on the logit model fitted to the n=1810 population (**Table 18**). This gave an expected proportion of 3.5%. This may be an overestimate due to the probabilities being higher than anticipated based on this model. In addition, this model which used a small proportion of the total population (n=1810) is being used to predict outcomes for a larger sample population. The subsample used to build the model was at higher risk which can have repercussions for extrapolating predictions at lower test results.

FIT Result (µg/g)	p(x) Probability of cancer/advanced adenoma for this test result*	Frequency of individuals with this FIT result	Total sample population with a result of <20 µg/g	f(x) Proportion of individuals with this FIT result out of the total number of those with a result of <20 µg/g	f(x)p(x)
0	0.0000	15404	24,930	0.6179	0.0000
1	0.0782	2718	24,930	0.1090	0.0085
2	0.0866	1739	24,930	0.0698	0.0060
3	0.0920	1098	24,930	0.0440	0.0041
4	0.0959	761	24,930	0.0305	0.0029
5	0.0991	552	24,930	0.0221	0.0022
6	0.1017	492	24,930	0.0197	0.0020
7	0.1040	330	24,930	0.0132	0.0014
8	0.1061	278	24,930	0.0112	0.0012
9	0.1079	225	24,930	0.0090	0.0010
10	0.1095	219	24,930	0.0088	0.0010
11	0.1110	148	24,930	0.0059	0.0007
12	0.1124	142	24,930	0.0057	0.0006
13	0.1137	226	24,930	0.0091	0.0010
14	0.1149	179	24,930	0.0072	0.0008
15	0.1161	129	24,930	0.0052	0.0006
16	0.1171	99	24,930	0.0040	0.0005
17	0.1181	75	24,930	0.0030	0.0004
18	0.1191	70	24,930	0.0028	0.0003
19	0.1200	46	24,930	0.0018	0.0002
*Probabilities based on the following logit model: $\text{logit}(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = -2.467 + 0.371 * \log(FIT\ result)$					**Σ f(x)p(x)**: 0.0354

Table 18: Estimation of the proportion of those with cancer or advanced adenoma with a FIT result of less than 20 µg/g. The logit model used to produce the estimated probabilities is given below the table. An assumption of a FIT result of 0 µg/g having a probability of 0 was made to obtain this estimation and discrete FIT values were used by rounding to the nearest whole number.

An assumption in this case was made that a result of '0' would provide a probability of zero for colorectal cancer/advanced adenomas whereas in reality this would not always be the case. For instance, there are adenomas which may bleed intermittently or have undetectable amounts of bleeding, leading to false negative results. In addition, to calculate an estimate, discrete FIT results were used by rounding to the nearest whole number; whereas data provided gave up to one decimal place. Based on these approximations, a NPV of 0.96 has been calculated for this test, which is similar to the values estimated and reported previously (**Table 19)**.

A potentially smaller ratio of false negative to true negative results based on the literature would be expected and so 0.96 could be considered an underestimate. The sensitivity of the test is also much lower than the other estimates due to this higher number of false negative results (0.39). Other methods could be used to estimate proportions for continuous test results. By repeating the analysis for Log(FIT+1) the proportion of people estimated with cancer/advanced adenoma with a FIT result below 20µg/g increases to 8.4%. This is due to the large number of individuals with a result of 0µg/g (15,404) which is a much greater overestimate.

This exploratory analysis suggests further models could be investigated which may fit the lower tail end of the data better to provide more accurate predictions for extrapolation.

	Positive Diagnostic Result	**Negative Diagnostic Result**	**Total**
FIT positive ≥ 20 µg/g	549	1261	1,810
FIT negative < 20 µg/g	874	24075	24,949
Total	1,423	25,336	26,759
Sensitivity: 0.39, Specificity: 0.95, PPV: 0.30, NPV: 0.96, Sample prevalence: 5.32%, FIT positivity 6.76%. **Excluded:** Those with positive test result (≥20 µg/g) but with no diagnostic outcome (27,066 – 307 = 26,759). This will likely underestimate the number of true positive results.			

Table 19: 2 by 2 table using the FIT data combined with logit model estimates for a FIT result of less than 20µg/g

4.0 DISCUSSION

4.1 Statement of principal findings

This study has demonstrated that including routinely available risk predictors in the screening algorithm alongside the FIT improved both model performance and test accuracy. At a threshold of 160µg Hb/g faeces, sensitivity improved from 30.78% to 33.15% using a risk adjusted model compared to FIT alone (setting the number of recalls the same/the same specificity). Furthermore, the risk-adjusted model for this sample population (threshold of 160µg Hb/g faeces) leads to the detection of 13 additional advanced adenomas and the same number of cancers (17 more high risk adenomas, 4 less intermediate risk adenomas) using FIT only as the comparator. Improvement was therefore seen in the number of high-risk adenomas detected by the risk-adjusted model.

Based on the results from these data, for every 1,000,000 people invited to screening, it can be estimated that 318 additional advanced adenomas (4,447/1,000,000) would be detected compared to FIT only (4,129/1,000,000). Although this approach would require external validation, the figures give the relative performance of this risk-based approach. The algorithm mainly improves detection in men compared to women.

A higher positive FIT rate (160µg/g) was independently associated with age, males, previous non-responders to screening, IMD score and sample return time ($p<0.05$). There was a negative association with the Southern compared to the Midlands hub ($p=0.04$) and ambient temperature was not significantly associated with a positive result ($p=0.47$). High temperature has been shown to be associated with reduced positivity.[52]

4.2 Strengths and weaknesses of the study

The main strength of the study was the quality of the data; the data were collected for the FIT pilot comparative study which was implemented within a live screening programme. In addition, routine data were used to develop the risk prediction model meaning no additional data collection was required, reducing costs and the burden on the screened participants. This approach also makes it feasible to implement risk-adjusted screening in practice by using data direct from the BCSS. The test thresholds analysed were those which were identified in the FIT pilot as having a similar positivity or referral rate to the gFOBT to

ensure adequate colonoscopy capacity. The TRIPOD guidelines were followed when developing and reporting the model to improve the quality of the study.

Limitations of the study include the lack of follow up data for participants with a result of <20 µg Hb/g faeces as interval cancers are not recorded on the BCSS. Ideally, follow up data for participants sent the FIT would be obtained from cancer registries (National Cancer Intelligence Network, or Office for National Statistics data linked through NHS number). A follow up period of two years would allow the clinical identification of existing cancers. Not all individuals had a diagnostic result if they cancelled or did not attend the appointment and this could cause potential selection bias if non-healthy participants tend to not have follow up colonoscopy. The pattern of attendance for diagnostic investigation seen in this study is however similar to that seen in the screening programme in general.[56] This approach can lead to partial verification bias and inflated test accuracy measures.[59 66] However, the results provided in this study give the relative performance of a risk-adjusted approach versus a regular screening approach. Multiple imputation can be used for missing predictor information but in the current study only 8 IMD scores were missing.

The selection bias in this study by only including those with a FIT result of ≥20 µg Hb/g faeces (as they have outcome data) has implications on making predictions for those with FIT results of less than 20 µg/g and subsequent model performance and test accuracy. The highly selected population affects external validation of the prediction model and therefore generalizability and accuracy. There are a greater proportion of participants at higher risk and with the outcome compared to a complete general risk screening population. Since the case mix is of higher risk in this model development study if applied to another population (with outcomes for all FIT results) the model would most likely need to be recalibrated. This can be achieved by adjusting regression coefficients by an adjustment factor e.g. calibration slope or by adjusting the intercept since there will be a difference in the sample prevalence from this model development study to a validation study using all cut-off ranges. The difference in case mix (in this case outcome prevalence) can affect the predictor-outcome associations and therefore when applied in a different sample this will affect model accuracy.[37 67]

The current model can estimate for FIT results less than 20 µg/g but the parameter estimates would be less precise as there is a greater degree of uncertainty around the

predicted probabilities for these individuals. Eligibility criteria and the study flow diagram are reported fully in this study to allow assessment of the applicability or generalizability of this model and its predictions in another dataset. Future research should investigate further predictors from the BCSS and obtain follow up data for those with a result <20 µg/g to improve generalizability.

Since not all participants have a reference standard in this study, this introduces partial verification bias. This is a common form of bias in screening studies as identified in the systematic review reported in **Chapter 2**, since the participant flow only tests those with a positive screening test result. Furthermore, the dataset used for this study uses routine data which is not specifically collected for research purposes. When assessing test accuracy characteristics, the sensitivity of the FIT and risk adjusted FIT would most likely be overestimated using these cut-offs (30 to 180) due to an underestimate of false negative results. The appropriate denominator to calculate sensitivity and specificity would ideally use all participants who adequately participated in FIT screening (27,066). By using a selected subsample of participants with results of ≥20 µg/g, presumably since these individuals are at higher risk of cancer/advanced adenomas being detected, the severity of disease detected would also be greater.[18 68] This may in turn inflate test accuracy parameters (higher sensitivity). In addition, by using this 'complete-case' approach, there is a reduction in sample size which can lead to reduced precision; bias is also introduced since not having a colonoscopy is associated with the risk of having colorectal cancer. As such the generalizability of these test accuracy measures can only be compared for those studies which assess test accuracy in those with a FIT result of ≥20 µg Hb/g.

Efforts were made in **Section 3.11** to estimate test accuracy measures for those with a result of <20 µg/g by comparing with literature estimates, data from **Chapter 5** and using a modelling approach to extrapolate the proportion of undiagnosed cancers/advanced adenomas. These approaches are crude estimates and should be compared to using follow up data and cancer registries for individuals with a negative result (although this too comes with differential verification bias due to different reference standards). Other methods to correct for partial and differential verification bias could be considered.[59]

Part of the increase in detection for the FIT in the pilot was due to increased uptake of this test compared with the gFOBT (66.4% vs 59.3%);[69] this study assumes the same uptake

seen with the pilot. In subsequent FIT screening rounds, there could however be a change in the uptake whereby non-responders to gFOBT are more likely to respond to the FIT, whereas non-responders to FIT may be less likely to respond to the next FIT. This could affect future detection rates and subsequently model performance. However, data from four rounds of a biennial FIT screening programme in the Netherlands showed that uptake increased from 60 to 63%, the same could be expected with this new test.[70] In addition, future iterations of the risk-based model could take into account more detailed previous screening episode factors which have been found to be predictive of uptake.[69]

As identified in the previous chapter for the systematic review, a common limitation of screening studies is the flow of participants due to uptake of the index test and then subsequent uptake of the reference standard. The final dataset includes those who adequately participated i.e. returned a FIT with a definitive positive or negative result and then those who subsequently went on to have the reference standard test. This can have two possible effects on the population. Individuals who participate in screening are likely to be at lower risk, due to the healthy screenee effect, compared to those who do not return a FIT. The same may be true for uptake of the reference standard but conversely this also includes the effect of elderly people and people with comorbidities who may not be recommended to undergo colonoscopy and who are often at higher risk. This can have an effect on the underlying case mix seen in the study.

4.3 Strengths and weaknesses in relation to other studies

Other studies which have investigated the added value of risk factors combined with the FIT have shown an improvement in model and test accuracy parameters as identified in the previous chapter. A study in the Netherlands developed a multivariable prediction model which combined the following risk factors: total calcium intake, family history, age and FIT result (OC-Sensor).[8] The risk-adjusted model had similar calibration and discrimination to the model developed in this study despite using different predictors. The AUC ROC improved from 0.69 to 0.76 when including the additional predictors compared with an improvement reported in this study of 0.63 to 0.66. The smaller increase seen in the current study compared to the model by Stegeman *et al*. is most likely due to the richer predictor data included in the model which came from a lifestyle questionnaire. The

questionnaire would however require a response and additional time from a participant whereas this study sought to use routine data.

In addition, sex of a participant was investigated in this study combining questionnaire data but was found to not be significant in the multivariable model. This suggests that the variance explained by sex was explained by another variable in the model by Stegeman et al.[71]. Previous studies have however suggested that male sex is associated with increased risk of CRC.[72 73 74] For instance, Kolligs *et al.* found that men had an odds ratio of 1.95 for advanced neoplasia (1.91-2.00) when compared with females.[74] This odds ratio is comparable to the odds ratio calculated for this study; adjusted 1.75 (1.413-2.17) and unadjusted 1.88 (1.53-2.31). Sex also has also been shown to affect the positivity of the test and as such is an important predictor to be included within the model.[75]

Stratification of risk using a logistic regression model combining age and sex with the FIT result has been investigated by Auge *et al*.[7] Colorectal cancer risk was stratified into 16 categories, these risk categories were then classified into 3 risk levels based on the positive predictive value. The authors suggest that this stratified approach could be used to prioritize higher risk individuals for colonoscopy examination. By categorizing risk however individual information is lost as the probabilities become standardized for all individuals in one group.[37] The current study gives an absolute risk prediction for each individual, providing a personalized and more accurate approach to screening. The discrimination of this model including routine predictors only was 0.68 (95% CI: 0.66-0.70) which is similar to what is reported in the current study 0.66 (95% CI: 0.63 - 0.69).

Deprivation has been shown to relate to the faecal haemoglobin concentration and has consequently been suggested to be included in risk scoring systems.[76] Although IMD score showed a decreased risk as the value increases (OR 0.997 95% CI: 0.99-1.00) this result was not significant in univariable analysis (p-value = 0.53) and was dropped from the multivariable model during stepwise selection. This study used data from two screening hubs, it may be the case that inclusion of data from other hubs across England could produce a significant result.

4.4 Practical Implications

This study utilised the data recorded routinely on the BCSS to develop a risk prediction model to ensure, as far as possible, that it could be implemented in practice without additional data collection. There are more data fields present on the screening system which could be investigated or combined for inclusion in the screening algorithm, particularly relating to screening history. The logistic regression risk adjusted model provides the absolute risk prediction (as a probability) for each individual and this can be used to make clinical decisions regarding screening referral by setting an appropriate 'risk threshold'.

Based on the results of this study, a risk-adjusted approach could be implemented at the point of screening to decide which participants are at greatest risk for more targeted colonoscopy referral. Further predictors exist on the BCSS which could be used to investigate a model which provides greater discrimination and corresponding test accuracy. This model would then need external validation in a new dataset to confirm study findings and to determine if similar performance is achieved. An impact study could then be used to assess using a risk-adjusted algorithm on screening outcomes.

4.5 Future research

Although the model improves when adding additional risk indicators to the screening algorithm, model performance metrics including Nagelkerke's R^2, AUC and the deviance suggest that the prediction of cancer/advanced adenomas at colonoscopy is not fully captured by the predictors used in the model. A similar AUC is achieved compared to models using routine data only identified in the systematic review reported in the previous chapter.[77] Future research could include the investigation of additional predictors from the BCSS database or obtaining richer data from questionnaires and electronic health records to improve predictive performance.

Additional predictors from the BCSS could include the introduction of the once only flexible sigmoidoscopy at age 55, since this will provide a protective effect for a number of years.[78] If an individual has had a previous negative colonoscopy this has also been shown to reduce risk.[79] Previous FIT results could be implemented in the algorithm and monitored over time as it has been shown that the Hb concentration relates to the detection of

adenomas and CRC in future screening rounds.[80] To make use of previous gFOBTs, the number of positive spots (or spot positivity %) could also be investigated whilst transitioning over to the FIT.[81] Lifestyle factors have been shown to have a significant effect on the risk of CRC (diet, alcohol, physical inactivity and being overweight).[82] Whilst the latter information is not currently included on the BCSS other sources such as electronic health records or questionnaires could be used to obtain this information.

The inclusion of more complex predictor information available from lifestyle questionnaires and from electronic health records may be better captured using machine learning algorithms. An alternative model to conventional logistic regression which could possibly perform better, might be a feed-forward artificial neural network (ANN). This model is highly flexible and, unlike logistic regression, does not require the strong assumption of linearity for combinations of variables and thus allows more complex nonlinear relationships to be included between predictors and the response variable in prediction models.[83]

Once the FIT is implemented in the NHS BCSP at the end of 2018, this risk-adjusted approach could be investigated in a similar way using the routine screening data once sufficient follow up data is available also enabling a more accurate risk positivity threshold to be derived. The algorithm led to greater detection in males compared to females which depending on screening programme aims will need greater investigation if a risk based approach is implemented in the future (e.g. using separate models for each sex). The acceptability of this difference to the population could also be examined. For instance, the screening programme may want to consider separate models for men and women. The detection rate in first time screenees also decreases with a risk-based approach for both males and females whilst doubles for male previous non-responders. Likewise the detection rate seen between responders/non responders/first time invitees will need consideration in future risk models by dissecting previous screening history in greater detail.

5.0 CONCLUSIONS

This research has investigated a risk-based approach to colorectal cancer screening by combining the FIT with routinely available predictors from the BCSS in a logistic regression model. This approach demonstrated an improvement in both model performance and test accuracy of the FIT when compared to using the FIT only. As the NHS BCSP prepares to transition to the FIT, these initial investigations have shown that further exploration of the BCSS for additional predictors which could be included in the algorithm may help to improve test accuracy and colonoscopy use.

Machine learning algorithms were identified as a potential avenue to explore in the previous chapter. Before embarking on further exploration of routine predictors, which may improve the discrimination of the risk adjusted model, a neural network will be investigated using the same dataset to determine if this improves model performance and test accuracy further without additional data collection. A neural network may allow more complex associations to be modelled and has been shown to have similar if not higher performance than logistic regression.[84-86] The use of more complex predictors and their associations from routine health records may be better captured with a machine learning algorithm.

6.0 REFERENCES

1. Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J, et al. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. Gut. 2016.
2. Lo SH, Halloran S, Snowball J, Seaman H, Wardle J, von Wagner C. Predictors of repeat participation in the NHS bowel cancer screening programme. British journal of cancer. 2015;112(1):199-206.
3. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 2015;136(5):E359-86.
4. Rees CJ, Bevan R, Zimmermann-Fraedrich K, Rutter MD, Rex D, Dekker E, et al. Expert opinions and scientific evidence for colonoscopy key performance indicators. Gut. 2016.
5. Moss S, Mathews C. NHS Bowel Cancer Screening Programmes: Evaluation of pilot of Faecal Immunochemical Test : Final report. National Screening Committee Website: Centre for Cancer Prevention, Wolfson Institute, Queen Mary University of London (QMUL); 2015.
6. Cooper JA, Moss SM, Smith S, Seaman HE, Taylor-Phillips S, Parsons N, et al. FIT for the future: a case for risk-based colorectal cancer screening using the faecal immunochemical test. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2016;18(7):650-3.
7. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014.
8. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
9. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. International journal of cancer. 2014;135(5):1203-12.
10. Omata F, Shintani A, Isozaki M, Masuda K, Fujita Y, Fukui T. Diagnostic performance of quantitative fecal immunochemical test and multivariate prediction model for colorectal neoplasms in asymptomatic individuals. European journal of gastroenterology & hepatology. 2011;23(11):1036-41.
11. Tao S, Haug U, Kuhn K, Brenner H. Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. British journal of cancer. 2012;106(8):1424-30.
12. Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, et al. A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. Gastrointestinal Endoscopy. 2015;81(3):719-27.
13. Otero-Estévez O, De Chiara L, Rodríguez-Berrocal FJ, Páez De La Cadena M, Cubiella J, Castro I, et al. Serum sCD26 for colorectal cancer screening in family-risk individuals: Comparison with faecal immunochemical test. British journal of cancer. 2015;112(2):375-81.
14. Nederlands Trial Register [Internet]. NTR Number NTR5874. Comparison study of two fecal immunochemical tests within a nationwide colorectal cancer screening program 27th May 2016 [cited 2016 17th November]. Available from: http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=5874.

15. Chiu HM, Ching JY, Wu KC, Rerknimitr R, Li J, Wu DC, et al. A Risk-Scoring System Combined With a Fecal Immunochemical Test Is Effective in Screening High-Risk Subjects for Early Colonoscopy to Detect Advanced Colorectal Neoplasms. Gastroenterology. 2016;150(3):617-25.e3.
16. Watson J, Shaw K, Macgregor M, Smith S, Halloran S, Patnick J, et al. Use of research questionnaires in the NHS Bowel Cancer Screening Programme in England: impact on screening uptake. Journal of medical screening. 2013;20(4):192-7.
17. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015;13:1.
18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of internal medicine. 2003;138(1):W1-12.
19. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer; 2009.
20. Department for Communities and Local Government. English indices of deprivation 2010 2011 [cited 2016 17th November]. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010.
21. Fraser CG, Allison JE, Halloran SP, Young GP. A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. Journal of the National Cancer Institute. 2012;104(11):810-4.
22. Department of Health. NHS public health functions agreement 2015-16. Service specification no.26 Bowel Cancer Screening Programmme 2014 [cited 2016 17th November]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/383200/1516_No26_NHS_Bowel_Cancer_Screening_Programme_Final.pdf.
23. NHS BCSP. Quality Assurance Guidelines for Colonoscopy. NHS BCSP Publication No 6 February 2011: NHS Cancer Screening Programmes; 2011 [cited 2016 17th November]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/427591/nhsbcsp06.pdf.
24. Cairns SR, Scholefield JH, Steele RJ, Dunlop MG, Thomas HJ, Evans GD, et al. Guidelines for colorectal cancer screening and surveillance in moderate and high risk groups (update from 2002). Gut. 2010;59(5):666-89.
25. Winawer SJ, Zauber AG. The advanced adenoma as the primary target of screening. Gastrointestinal endoscopy clinics of North America. 2002;12(1):1-9, v.
26. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. Gut. 2007;56(11):1585-9.
27. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
28. Hadley Wickham and Romain Francois. dplyr: A Grammar of Data Manipulation. R package version 0.4.1 ed2015.
29. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. European journal of epidemiology. 2011;26(4):261-4.
30. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12:77.
31. Venables WNR, B. D. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

32. Beck M. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. R package version 1.3.1 ed2015.
33. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009.
34. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-13 ed2014.
35. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of clinical epidemiology. 1996;49(8):907-16.
36. Field A. Discovering Statistics Using R. London: SAGE Publications Ltd; 2012.
37. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015;162(1):W1-73.
38. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology. 2001;54(8):774-81.
39. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions Journal of the Royal Statistical Society Series B (Methodological). 1974;36(2):111-47.
40. Altman DG. Problems in dichotomizing continuous variables. American journal of epidemiology. 1994;139(4):442-5.
41. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ : British Medical Journal. 2006;332(7549):1080-.
42. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.
43. McDonald PJ, Strachan JA, Digby J, Steele RJ, Fraser CG. Faecal haemoglobin concentrations by gender and age: implications for population-based screening for colorectal cancer. Clinical chemistry and laboratory medicine : CCLM / FESCC. 2012;50(5):935-40.
44. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 2010;21(1):128-38.
45. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. . Communications in Statistics. 1980;A10:1043-69.
46. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Statistics in medicine. 2013;32(1):67-80.
47. Nagelkerke N. A note on a general definition of the coefficient of determination. Biometrika. 1991(78):691–2.
48. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the Predictiveness of a Marker with Its Performance as a Classifier. American journal of epidemiology. 2008;167(3):362-8.
49. van Rossum LG, van Rijn AF, van Oijen MG, Fockens P, Laheij RJ, Verbeek AL, et al. False negative fecal occult blood tests due to delayed sample return in colorectal cancer screening. International journal of cancer. 2009;125(4):746-50.
50. van Roon A, van Dam L, Zauber A, van Ballegooijen M, Borsboom G, Steyerberg E, et al. Guaiac-based faecal occult blood tests versus faecal immunochemical tests for colorectal cancer screening in average-risk individuals (Protocol). Cochrane Database of Systematic Reviews 2011. 2011(8):Art. No.: CD009276. DOI: 10.1002/14651858.CD009276.
51. Daly JM, Bay CP, Xu Y, Levy BT. Effect of Ambient Temperature Variations on Positivity of Manual Fecal Immunochemical Tests. Journal of primary care & community health. 2015;6(4):243-9.

52. Symonds EL, Osborne JM, Cole SR, Bampton PA, Fraser RJ, Young GP. Factors affecting faecal immunochemical test positive rates: demographic, pathological, behavioural and environmental variables. Journal of medical screening. 2015.
53. Grazzini G, Ventura L, Zappa M, Ciatto S, Confortini M, Rapi S, et al. Influence of seasonal variations in ambient temperatures on performance of immunochemical faecal occult blood test for colorectal cancer screening: observational study from the Florence district. Gut. 2010;59(11):1511-5.
54. UK Met Office. UK Met Office Weather Open Data 2016 [cited 2016 17th November 2016]. Available from: https://datamarket.azure.com/dataset/explore/datagovuk/metofficeweatheropendata) https://data.gov.uk/metoffice-data-archive.
55. Wieten E, Schreuders EH, Grobbee EJ, Nieboer D, Bramer WM, Lansdorp-Vogelaar I, et al. Incidence of faecal occult blood test interval cancers in population-based colorectal cancer screening: a systematic review and meta-analysis. Gut. 2018.
56. Logan RFA, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. Gut. 2011.
57. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. Journal of clinical epidemiology. 1995;48(12):1495-501.
58. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of clinical epidemiology. 1995;48(12):1503-10.
59. de Groot JA, Bossuyt PM, Reitsma JB, Rutjes AW, Dendukuri N, Janssen KJ, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ (Clinical research ed). 2011;343:d4770.
60. Raffle A, Gray M. Screening Evidence and Practice. New York, United States: Oxford University Press; 2007.
61. Sanduleanu S, le Clercq CM, Dekker E, Meijer GA, Rabeneck L, Rutter MD, et al. Definition and taxonomy of interval colorectal cancers: a proposal for standardising nomenclature. Gut. 2015;64(8):1257-67.
62. Lee JK, Liles EG, Bent S, Levin TR, Corley DA. Accuracy of fecal immunochemical tests for colorectal cancer: systematic review and meta-analysis. Annals of internal medicine. 2014;160(3):171-.
63. Office for National Statistics. Cancer Registration statistics, England: 2016: ONS; 2016 [Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/final2016.
64. Digby J, Fraser CG, Carey FA, Lang J, Stanners G, Steele RJ. Interval cancers using a quantitative faecal immunochemical test (FIT) for haemoglobin when colonoscopy capacity is limited. Journal of medical screening. 2016;23(3):130-4.
65. Launois R, Le Moine JG, Uzzan B, Fiestas Navarrete LI, Benamouzig R. Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. European journal of gastroenterology & hepatology. 2014;26(9):978-89.
66. Naaktgeboren CA, de Groot JAH, Rutjes AWS, Bossuyt PMM, Reitsma JB, Moons KGM. Anticipating missing reference standard data when planning diagnostic accuracy studies. BMJ : British Medical Journal. 2016;352:i402.
67. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American journal of epidemiology. 2010;172(8):971-80.

68. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England journal of medicine. 1978;299(17):926-30.
69. Moss S, Mathews C, Day TJ, Smith S, Halloran SP. A faecal immunochemical test for haemoglobin (FIT) markedly increased participation in a colorectal cancer screening pilot in England. Poster session presented at: Third NAEDI research conference March 26-27 2015. 2015.
70. van der Vlugt M, Grobbee EJ, Bossuyt PMM, Bongers E, Spijker W, Kuipers EJ, et al. Adherence to colorectal cancer screening: four rounds of faecal immunochemical test-based screening. Br J Cancer. 2017;116(1):44-9.
71. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
72. Nguyen SP, Bent S, Chen YH, Terdiman JP. Gender as a risk factor for advanced neoplasia and colorectal cancer: a systematic review and meta-analysis. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2009;7(6):676-81.e1-3.
73. Brenner H, Hoffmeister M, Arndt V, Haug U. Gender differences in colorectal cancer: implications for age at initiation of screening. British journal of cancer. 2007;96(5):828-31.
74. Kolligs FT, Crispin A, Munte A, Wagner A, Mansmann U, Goke B. Risk of advanced colorectal neoplasia according to age and gender. PloS one. 2011;6(5):e20076.
75. Symonds EL, Osborne J, Cole SR, Bampton P, Fraser R, Young GP. Gender differences in faecal haemoglobin concentration. Journal of medical screening. 2016;23(1):54.
76. Digby J, McDonald PJ, Strachan JA, Libby G, Steele RJ, Fraser CG. Deprivation and faecal haemoglobin: implications for bowel cancer screening. Journal of medical screening. 2014;21(2):95-7.
77. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014;147(3):628-+.
78. Geurts SM, Massat NJ, Duffy SW. Likely effect of adding flexible sigmoidoscopy to the English NHS Bowel Cancer Screening Programme: impact on colorectal cancer cases and deaths. British journal of cancer. 2015;113(1):142-9.
79. Brenner H, Haug U, Arndt V, Stegmaier C, Altenhofen L, Hoffmeister M. Low risk of colorectal cancer and advanced adenomas more than 10 years after negative colonoscopy. Gastroenterology. 2010;138(3):870-6.
80. Digby J, Fraser CG, Carey FA, Diament RH, Balsitis M, Steele RJ. Faecal haemoglobin concentration is related to detection of advanced colorectal neoplasia in the next screening round. Journal of medical screening. 2016.
81. Geraghty J, Butler P, Seaman H, Snowball J, Sarkar S, Blanks R, et al. Optimising faecal occult blood screening:retrospective analysis of NHS Bowel Cancer Screening data to improve the screening algorithm. British journal of cancer. 2014;111(11):2156-62.
82. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. British journal of cancer. 2011;105(S2):S77-S81.
83. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology. 1996;49(11):1225-31.
84. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8 Suppl):1636-42.

85. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.
86. Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular Cancer. 2005;4:29-.

7.0 APPENDICES

Appendix 1: Ethical and Research Approval Letters

Bowel Cancer Screening Programme
Research Committee

John Scholefield
Chair

C/o NHS Cancer Screening Programmes
Fulwood House
Old Fulwood Road
Sheffield
S10 3TH

Tel: 0114 2013040

rachel.athorn@phe.gov.uk

Jennifer Cooper
Doctoral Student
Division of Health Science
Warwick Medical School
University of Warwick
Coventry
CV4 7AL

Jennifer.Cooper@warwick.ac.uk

17th July 2015

Dear Jennifer Cooper,

The Bowel Cancer Screening Programme (BCSP) Research Committee met on 15th July 2015 to discuss your research plans: *Risk-Adjusted Colorectal Cancer Screening Using the FIT: Development of a Risk Prediction Model. ID 152*

The Committee gave their support to the project.

As a condition of support, the BCSP Research Committee requires you to keep them informed of developments with the project, including any changes of status, any significant adverse events, when completed, and when written up.

Please note that any applications requiring patient identifiable data from the BCSP programme will also require PHE ODR (Office of Data Release) approval.

The BCSP Research Committee requires you to notify them promptly of any incidents that would be recorded on the National Research Ethics Service (NRES) Breaches Register. Undertaking research within the Screening Programme following receipt of this letter of support assumes your agreement to fulfil this obligation. NRES has the potential to share information with the BCSP Research Committee regarding any breaches of ethics related to projects involving the BCSP.

The Committee wishes you well with your project.

Yours sincerely

Rachel Athorn MSc BMedSci
On behalf of the NHS BCSP Research Committee.

23rd June 2015

Warwick Medical School

PRIVATE
Miss Jennifer Cooper
PhD Student
Health Sciences
Warwick Medical School
University of Warwick
Coventry
CV4 7AL

Dear Miss Cooper,

Study Title and BSREC Reference: *Risk-Adjusted Colorectal Cancer Screening using the FIT: Development of a Risk Prediction Model* REGO-2015-1575

Thank you for submitting the above-named project to the University of Warwick Biomedical and Scientific Research Ethics Committee for research ethical review.

I am pleased to advise that research ethical approval is granted.

May I take this opportunity to wish you success with the study, and to remind you that any substantial amendments require approval from BSREC before they can be implemented. Please keep a copy of the original signed version of this letter with your study documentation.

Yours sincerely

P.P. Graham Hewitt

Professor Scott Weich
Chair
Biomedical and Scientific
Research Ethics Sub-Committee

Biomedical and Scientific Research Ethics Sub-Committee
A010 Medical School Building
Warwick Medical School,
Coventry, CV4 7AL.
Tel: 02476-528207
Email: BSREC@Warwick.ac.uk

Medical School Building
The University of Warwick
Coventry CV4 7AL United Kingdom
Tel: +44 (0)24 7657 4880
Fax: +44 (0)24 7652 8375

THE UNIVERSITY OF WARWICK

Appendix 2: System Level Security Policy for this Research

System Level Security Policy

Organisation Name: University of Warwick

System Details

1. The System shall be known as RAFS (Risk Adjusted FIT Study)

2. The System's responsible owner shall be Jennifer Cooper (lead researcher for the study)

3. The System's Caldicott Guardian or Data Controller shall be Professor Aileen Clarke, Head of Health Sciences Division at Warwick Medical School. The University is not required to have a Caldicott guardian.

System Security

4. Security of the system shall be governed by the corporate security policy of the University of Warwick specified by the regulations, policies, practices and guidance within the Information Security Framework. Please see the following webpage for an overview of the Information Security Framework:
http://www2.warwick.ac.uk/services/gov/informationsecurity/nav/

5. The System's responsible security manager shall be Julie Sheriff Warwick Medical School IT Team Leader

6. The security manager duties shall include:

- Ensuring that those individuals given access to the data (the lead researcher Jennifer Cooper and associated PhD supervisors Dr. Sian Taylor-Phillips and Dr. Nick Parsons) comply with this SLSP
- This includes **(i)** Ensuring all procedures follow the University of Warwick's Information Security Framework **(ii)** Ensuring the research follows the University of Warwick's 'Research Code of Practice', which links to many other policies and regulations including Data Protection Guidelines, Freedom of Information Guidelines and encompasses the University's Research Data Management Policy **(iii)** Ensuring that the research follows the Warwick Medical School Information Security Policy on the use of Portable devices.
- Please see following link for the full Research Code of Practice http://www2.warwick.ac.uk/services/ris/research_integrity/code_of_practice_and_policies/research_code_of_practice/
- The security manager will also be responsible for setting up a secure encrypted desktop computer and encrypted folder to store the data on a secure server within a locked room
- Carrying out a system risk assessment on an annual basis

7. The System shall incorporate the following security countermeasures

Physical security measures (E.g. secure room, cabinet, etc)
The data will be accessed on one desktop computer which will be located in a locked room requiring an access code only known to specific individuals at Warwick Medical School. Data will be stored in an encrypted folder on a secure server and will only be accessed from an encrypted computer at the University of Warwick. The encrypted folder will house the raw data as well as the data analysis files. File servers, failover servers and backups are managed by the University of Warwick's Server Provisioning team and kept in a University of Warwick data centre. Data centres have a strictly policed access control system and CCTV surveillance. Occasional

visitors/contractors are always escorted by an appropriate member of staff. The Medical School building also requires University ID card access so only authorised individuals can enter the building. No paper files will be used for this research study.

Logical measures for access control and privilege management
The electronic data will be stored within an encrypted folder on the University of Warwick's file server and will only be accessed from an encrypted computer in a locked room requiring an access code within the medical school building. The encrypted computer requires a unique username and password to log onto the computer. Access to the encrypted folder will be restricted to the lead researcher (Jennifer Cooper) and associated PhD supervisors (Dr. Sian Taylor-Phillips and Dr. Nick Parsons).

Network security measures (E.g. firewalls, network segregation, etc)
The University of Warwick's network is managed and monitored by the University's IT Services Network and Security Services team. The University's network is protected by Cisco designed infrastructure incorporating firewalls, VLANS (used to secure distinct areas of the University's campus) and Intrusion Detection Software to alert the Network Services Team of potential risks. The Security team also uses a comprehensive system for logging and monitoring network activity. The ITS Security team can undertake vulnerability assessments and penetration testing. This will be requested from the IT team on an annual basis to check the security of the data.

System Management

8. The System shall be developed / provided by Jennifer Cooper (Lead Researcher) at Warwick with the assistance of the Medical School IT Services team and the Administration and Governance team. Jennifer Cooper has received Information Security training at Warwick University, which covered the key principles of information security outlined in the University of Warwick Information Security Framework. Jennifer Cooper has also completed an online information security essentials course available to staff and students at Warwick.

9. The System shall be implemented and maintained by Jennifer Cooper (the lead researcher) and associated supervisors (Dr. Sian Taylor-Phillips and Dr. Nick Parsons). Jennifer Cooper will be responsible for receiving the FIT pilot data securely from the ODR (Office for Data Release) and maintaining the database on a secure University of Warwick IT server. ODR can supply the data in an Excel or CSV file format on an encrypted disk. Statistical packages will be used to analyse the data; R Studio and Stata for Windows. The Medical School IT team will ensure that the data and associated files are stored within an encrypted folder on a secure server and that it is only accessed from an encrypted desktop PC.

IT Services will automatically backup the encrypted data and backups will therefore be encrypted. At the end of the study and after associated publications, IT services will be notified for secure disposal of the information and data.

10. The System shall be shared or used by the following organisations.....

Public Health England control Bowel Cancer Screening Programme data, including the FIT pilot data. The Office for Data Release (ODR) will manage the release of data from PHE to the lead researcher (Jennifer Cooper). The University of Warwick will use this SLSP and no data will be shared with any other organisations.

System Design

11. The System shall comprise

A single encrypted desktop computer will be used to access and process the data. This computer will be located within a locked room requiring an access code within the medical school building, which is ID card access only. An individual username and strong password will be required to login onto the PC.

The computer is a fully managed Windows 7 computers setup to a secure University standard by the IT Services department (anti-virus protection, firewall, security and software updates managed by IT Services, no administrative access for users, etc). The computer will be locked if the researcher leaves the desk and securely logged off at the end of each day.

The hard disk of this desktop computer will be encrypted and an encrypted folder will be set up with help of the IT services team which will be stored on a secure University file server. This secure file server is in a secure data centre at the University of Warwick where only specific authorised personnel have access using a Warwick ID staff card.

The data will be received by the lead researcher (Jennifer Cooper) from the ODR on an encrypted disk. This data will then be stored securely within the encrypted folder on the University's secure file server. Data analysis will be done solely on the encrypted PC as detailed above.

The main outcome of interest is whether the accuracy of the FIT (faecal immunochemical Test) for colorectal cancer improves when combining peoples risk factors into a risk prediction model compared to a model which uses just the FIT alone. All individual data will be combined to develop the models so no individual data will be reported. In addition no identifiable information will be used for analysis for example IMD score is used instead of postcode and age instead of date of birth will be used. Only fields essential for data analysis and the study will be requested from the ODR, any de-identified data or data not required will be removed from the database.

The following data-fields are contained within the FIT pilot data and will be used for data analysis:

Data Requested For Analysis

Original Data Requested		
FIT Evaluation Data Extract - Colonoscopy Assessment EPISODE_ID (pseudonymised) APPOINTMENT_FIRST_OFFERED_DATE APPOINTMENT_DATE FIT_FOR_COLONOSCOPY FIT_FOR_RADIOLOGICAL_TESTS	**FIT Evaluation Data Extract - Diagnostic Tests** EPISODE_ID (pseudonymised) DIAGNOSTIC_TEST_ID FIRST_OFFERED_DATE ATTENDED_DATE TEST_TYPE INTENDED_EXTENT ACTUAL_EXTENT DIAGNOSTIC_TEST_RESULT DIAGNOSTIC_TEST_OUTCOME REASON_ANOTHER_REQUIRED FIT_FOR_COLONOSCOPY	**FIT Evaluation Data Extract - Episodes** EPISODE_ID EPISODE_STATUS EPISODE_SUBTYPE PREVALENT_INCIDENT EPISODE_SEQUENCE_NUMBER INVITED SENT_TEST_KIT RETURNED_KIT ADEQUATELY_PARTICIPATED DEFINITIVE_POSITIVE EPISODE_RESULT AGE_AT_EPISODE_START
FIT Evaluation Data Extract - Polyps POLYP_ID (pseudonymised) DIAGNOSTIC_TEST_ID POLYP_SIZE LOCATION CLASS REMOVAL_TYPE TYPE ARCHITECTURE	**FIT Evaluation Data Extract - Subjects** SCREENING_SUBJECT_ID GENDER IMD_SCORE RANK_OF_IMD_SCORE AGE_AT_EXTRACT	**FIT Evaluation Data Extract - Test Kits** TEST_KIT_ID (pseudoymised) EPISODE_ID (pseudonymised) SUBJECT_ID (pseudonymised) TEST_KIT_TYPE DATE_KIT_SENT G_SAMPLE_DATE_1 SAMPLE_1_SPOT_1

03/03/2016 Version 3.0

CARCINOMA HISTOLOGY_REPORT_DATE		SAMPLE_1_SPOT_2 G_SAMPLE_DATE_2 SAMPLE_2_SPOT_1 SAMPLE_2_SPOT_2 G_SAMPLE_DATE_3 SAMPLE_3_SPOT_1 SAMPLE_3_SPOT_2 F_SAMPLE_DATE DATE_KIT_LOGGED G_DATE_KIT_READ F_DATE_KIT_READ KIT_RESULT F_FIT_KIT_RESULT HUB_CODE HUB_NAME

Additional Data Requested

Polyps:
Polyp intervention Modality (i.e. Mucosal Resection (EMR), Polypectomy or Submucosal Dissection (ESD))
Secondary piece (Yes or No)
Excised (Yes or No)
Retrieved (Yes or No)
Pathology Lost (Yes or No)
Reason pathology lost (Destroyed During Processing or Lost in Transit)
Size (from Histology)

Diagnostic Tests:
Screening Centre where the test took place

Subjects:
Subject hub
Subject screening centre

Episodes:
SUBJECT_ID
Date a reminder letter was sent
Latest event status

Other tables.

Pre-invitations:
Subject ID
Episode ID
Hub Code
Pre_Invitation_Date
Invitation_type (FIT / gFOBt)

Cancers:
Cancer ID
Episode ID
Size
Type
Location
Excision type
Primary procedure
Final pre-treat T stage
Final pre-treat N stage
Final pre-treat M stage
Pathological Dukes staging

Previous Episodes:

Subject ID
Episode ID
Episode type
Start date of the Episode
Episode sub type
Prevalent / incident
Episode sequence number
Invited?
Sent test kit?
Returned a kit?
Adequately participated?
Test kit definitive positive?
Episode result & description

03/03/2016 Version 3.0

Previous Episodes - Test Kits

TEST_KIT_ID (pseudoymised)
EPISODE_ID (pseudonymised)
SUBJECT_ID (pseudonymised)
TEST_KIT_TYPE
DATE_KIT_SENT
G_SAMPLE_DATE_1
SAMPLE_1_SPOT_1
SAMPLE_1_SPOT_2
G_SAMPLE_DATE_2
SAMPLE_2_SPOT_1
SAMPLE_2_SPOT_2
G_SAMPLE_DATE_3
SAMPLE_3_SPOT_1
SAMPLE_3_SPOT_2
F_SAMPLE_DATE
DATE_KIT_LOGGED
G_DATE_KIT_READ
F_DATE_KIT_READ
KIT_RESULT
F_FIT_KIT_RESULT
HUB_CODE
HUB_NAME

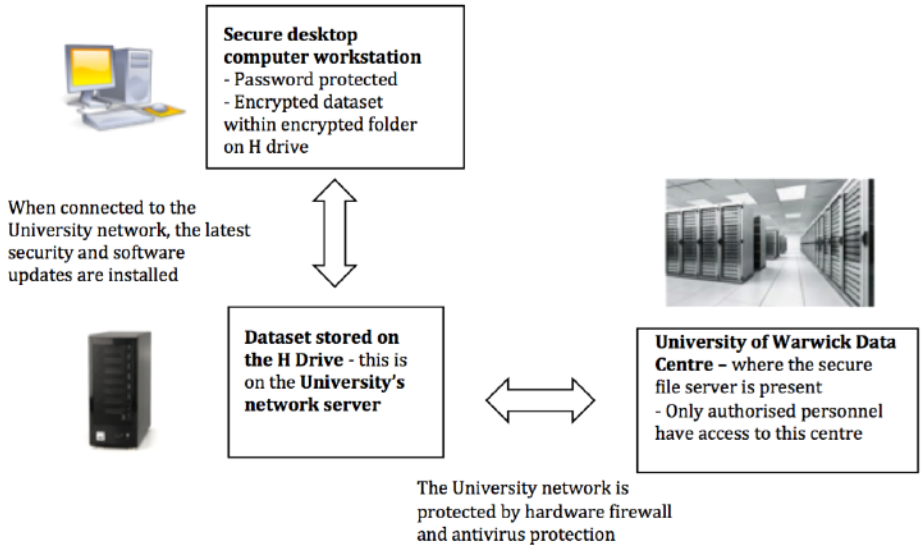


***Figure 1**: Electronic System Diagram*

Operational Processes

12. The patient identifiable / sensitive data will be collected .....

The data for this PhD project has previously been collected for the UK FIT pilot study which is assessing whether this new screening test should replace the current screening test. A research team based at Queen Mary University of London is carrying out the analysis of the FIT pilot study. The HSCIC (Health and Social Care Information Centre) own the FIT pilot data and this will be

03/03/2016 Version 3.0

provided by the Office for Data Release (ODR) to the lead researcher (Jennifer Cooper). The data will be transferred by the ODR using an encrypted disk and the data will be encrypted where it is stored and processed. The Bowel Cancer Screening Programme Research Committee (BCSP RC) has also approved this study ID 152.

13. The data will be stored .....

No paper records of the data will be used. The ODR (Office for Data Release) will provide the data using an encrypted disk. This data will then be encrypted and transferred to an encrypted folder on the University of Warwick file server. The server will be backed up automatically by IT Services and both the primary data on the server and the backup data will be encrypted. Only University staff with unique login credentials can access this server. Symantec Encryption Software will be used to encrypt the PC hard disk as well as files and folders to using AES 256 encryption. Access to the encrypted folder will be limited to Jennifer Cooper (lead researcher) and associated PhD supervisors (Dr. Sian Taylor- Phillips and Dr. Nick Parsons). A single secure desktop computer will be used to access the data in a locked room requiring a code for entry. The desktop computer is protected using an individual login and password and the whole disk is encrypted.

The FIT pilot data is de-identified and contains the data fields as shown in the table above. No identifiable information will be used for analysis, for example IMD score is used instead of postcode and age instead of date of birth will be used. All individual data will be combined to develop the models so no individual data will be reported. In addition, only fields essential for data analysis and the study will be requested from the ODR, any data not required will be removed from the database. Encryption is being used to ensure data is protected as some fields could be considered sensitive and full dates (such as appointment dates) are being provided within the dataset.

14. The data will be processed .....

A single encrypted desktop computer will be used to access and process the data. This computer is located in a secure room requiring a code within the Medical School building. A unique login and password is required to log onto the computer and University network. The data will be stored on the University of Warwick file server as opposed to the individual machine. The data will be encrypted and stored within an encrypted folder using Symantec Encryption software to the AES 256 standard. Statistical packages will be used to analyse the data (R Studio and Stata for Windows) and the associated files will be stored within the encrypted folder. No further copies of the data will be made (except for the automated backups of the encrypted data). University policy (the University Information Security Framework and Medical School policy on use of Portable devices) does not permit sensitive data to be stored in an unencrypted format on any portable device or removable device.

No personal or sensitive data will be printed out for this study. The University has an 'Information Classification and Handling Procedure' which is part of the University's Information Security Framework, this will also be followed where appropriate. This classifies information by its confidentiality, criticality and value and gives guidance on how to protect both electronic and paper information.

The following link gives further information on the Information Classification and Handling Procedure - http://www2.warwick.ac.uk/services/gov/informationsecurity/handling/

15. The System's authorised users shall be .....

Jennifer Cooper (the lead researcher) and the supervisors of the PhD project (Dr. Sian Taylor-Phillips and Dr. Nick Parsons) will follow this SLSP. The system data will only be stored at the

University of Warwick.

16. When the system or its data has completed its purpose / has become redundant or is no longer needed, the following methods will be adopted to dispose of equipment, back-up media or other stored data .....

The University's IT team will be informed when the study and data has completed its purpose so that the data can be disposed of securely. ITS (IT services) can digitally shred the encrypted data and can destroy any encryption keys so that the data cannot be recovered. All computer hardware will be disposed via the secure IT Services disposal service.

As per the Data Protection Act, data will be 'kept for no longer than is absolutely necessary'. Since this study is contributing to a PhD project, this will be for the duration of the project (expected hand in date: February 2018) with the possibility of extension if agreed and if there is a legitimate reason.

System Audit

17. The System shall benefit from the following internal / external audit arrangements *(Please list all arrangements)* The ITS Security team at the University can undertake vulnerability assessments and penetration testing of the system to check the security of the data. This will be requested on an annual basis and vulnerability scans can be carried out more frequently as required. The ITS security service can provide scan reports and these will be passed onto the system owner (Jennifer Cooper) and the Information Security team to assess and mitigate any vulnerabilities.

18. The System shall be risk assessed every 12 months

18.1 - During vulnerability assessments, the ITS Security Service classifies security issues as 4 different risk levels which are based on the Common Vulnerability Scoring System (CVSSv2) which is used to rate IT vulnerability and determine the level of response.

18.2 - A risk management / security improvement plan shall be established to address all unacceptable risks. For instance, all potential risks to data will be noted and described in a spreadsheet and a risk log kept throughout the study.

System Protection

19. The System shall benefit from the following resilience / contingency / disaster recovery arrangements.....

Both client computer and server computers are configured and fully managed by IT Services. Data will be saved to an encrypted folder on an IT Services managed server. IT Services managed servers are backed up daily to a tape library, keeping data for up to 2 years (7 dailies, 5 weeklies, 12 monthlies, 2 yearlies). IT Services provide a data recovery service and this is routinely tested. Servers are hosted across multiple data centres on clustered hardware. In the event of a power failure, the data centre's batteries or separate diesel generator become operational. The encryption service is administered by the Security team within IT Services and in case of lost encryption keys, there is a recovery service.

20. In the event of serious disruption or total system failure, business continuity shall be provided by the following means .....

Ensuring business continuity is a key part of the University's Information Security Framework and there is a University of Warwick Emergency Planning Policy to ensure critical business systems

continue to run effectively and to ensure safety of University members.

The University of Warwick Information Security Framework -

http://www2.warwick.ac.uk/services/gov/informationsecurity/nav/

The University of Warwick Emergency Planning Policy -

http://www2.warwick.ac.uk/services/gov/emerg-planning/

If the client computer fails another computer in the same room could be encrypted. If the server fails then IT Services would migrate the encrypted data to another server.

21. In the event of a security or confidentiality breach occurring the following procedure shall be followed .....

If there is any concern of a confidentiality or security breach, the Warwick Information Security Team will be informed as well as the lead researchers line managers (in this case the Supervisors of the PhD project; Dr. Sian Taylor-Phillips and Dr. Nick Parsons), the Medical School IT Manager, the Head of Department (The Dean of Warwick Medical School Professor Sudhesh Kumar and the Head of Division of Health Sciences Professor Aileen Clarke) and the Deputy Registrar's Office. There is a set procedure to follow in instances such as these which is described in detail within the Warwick Information Security Framework.

http://www2.warwick.ac.uk/services/gov/informationsecurity/faqs/incidents

System Level Security Policy Ownership

22. This SLSP shall be the responsibility of Jennifer Cooper (lead researcher for the project) and authorised PhD supervisors (Dr. Sian Taylor-Phillips and Dr. Nick Parsons) with assistance of the WMS (Warwick Medical School IT team), the Warwick University Information Security Team and the WMS administration and governance department.

22.1 – This SLSP will be reviewed on an annual basis for its completeness and for any relevant updates in the University of Warwick Security Policy, Research Code of Practice (which encompasses the University's Research Data Management Policy) and legislation such as Data Protection Guidelines and Freedom of Information Guidelines.

23. The SLSP shall be available / distributed to ..... All authorised system users (Jennifer Cooper and Supervisors to the project Dr. Sian Taylor-Phillips and Dr. Nick Parsons), The Office for Data Release (ODR) and The Warwick Medical School IT team and the IT Services Security team.

Data Protection Registration

24. Please confirm that your organisation has Data Protection Registration to cover the purposes of analysis and for the classes of data requested.

The University of Warwick's Data Protection Registration Number is **Z5856740**

Date Registered: 07 November 2001
Registration Expires: 06 November 2016
Data Controller: University Of Warwick

03/03/2016 Version 3.0

Address:
Deputy Registrar's Office
University House
University Of Warwick
Coventry
CV4 8UW

03/03/2016 Version 3.0

Appendix 3: R scripts used for model development and to assess performance

```
#Logistic Regression Modelling#
# ------------------------------------------------------------------------------------------------------------
# Read Data in for modelling
ccFIT20 <- read.csv("FITroutine.csv")

FIT <- read.csv("FIT.AllRecords.csv")

# ------------------------------------------------------------------------------------------------------------
#Developing Risk Prediction Model - Stepwise Regression and Cross-Validation

#Load the Mass library
library(MASS)

#Load cross validation functions (see bottom of script)

source("CVfunction.R")
ccFIT20 <- data.frame(ccFIT20, CANCERAA = ccFIT20$Binary.outcome)

#Summarise a baseline model
base.mod <- glm(CANCERAA ~ 1, data = ccFIT20 ,family=binomial(link="logit"))
summary(base.mod)

# ------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------
#Model Fitting
#Backwards Elimination Stepwise

###1 Start with full model

cclog.mod.3 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + IMD_SCORE +
prev.incident, data = ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "LRT")

#suggests IMD_SCORE should be removed, run cross validation

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)

#Gives coefficient estimates and associated confidence intervals
exp(cclog.mod.3$coefficients)
exp(confint(cclog.mod.3))

# ------------------------------------------------------------------------------------------------------------
###2 Removing IMD as a variable as least significant as signified by p-value

cclog.mod.3 <- glm(Binary.outcome ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident,
data = ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "Chisq")

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
###3 Remove age

cclog.mod.3 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + GENDER + prev.incident, data = ccFIT20,
family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "Chisq")

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)
```

```
#All remaining variables significant

# ------------------------------------------------------------------------------------------------------------
#For completeness:
###4 Remove previous screening history (prev.incident)

cclog.mod.3 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + GENDER, data = ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "Chisq")

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
###5 Remove Gender (sex)

cclog.mod.3 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1), data = ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "Chisq")

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
###6 Remove FIT

cclog.mod.3 <- glm(CANCERAA ~ 1, data = ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.3)

drop1(cclog.mod.3, test = "Chisq")

cv.logreg(mod = cclog.mod.3, nreps = 10, seedno = 253636)



#Previous screening history, gender and FIT only to be retained.
#Force age back into model due to clinical significance.
# ------------------------------------------------------------------------------------------------------------
#########Final Model#########

#Force age back into the model due to clinical signficance, systematic review and literature associations

cclog.mod.2 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident, data
= ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.2)
exp(cclog.mod.2$coefficients)

# ------------------------------------------------------------------------------------------------------------
#Investigating possible interactions
# ------------------------------------------------------------------------------------------------------------
# Interactions
#1 Gender and Age Interaction
cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
(AGE_AT_EPISODE_START*GENDER), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)

#Gender and Age Interaction not significant - could cross validate/bootstrap to see if it changes the result

source("CVANNFunctions.R")

cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
```

```
#2.          Gender + FIT

cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
((log(TRANS_FIT_KIT_RESULT + 1))*GENDER), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)


cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
#3.          GENDER+ prev.incident
cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
(prev.incident*GENDER), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)


cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------


#4.          FIT+age

cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
((log(TRANS_FIT_KIT_RESULT + 1))*AGE_AT_EPISODE_START), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)


cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)
# ------------------------------------------------------------------------------------------------------------
#5.          FIT+ prev.incident

cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
((log(TRANS_FIT_KIT_RESULT + 1))*prev.incident), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)


cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)
# ------------------------------------------------------------------------------------------------------------
#6.          AGE + prev.incident

cclog.mod.I <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident +
(AGE_AT_EPISODE_START*prev.incident), data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.I)
exp(cclog.mod.I$coefficients)


cv.logreg(mod = cclog.mod.I, nreps = 10, seedno = 253636)

###No interactions significant at the p=0.05 significance level
# ------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------
#Cross-validation
#Cross validate the model for final cross validated deviance measure

cv.logreg(mod = cclog.mod.2, nreps = 10, seedno = 253636)

# ------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------
```

```
#For reporting bootstrapped coefficient bias and confidence intervals

library("boot")

logit.bootstrap <- function(data, indices) {

  d <- data[indices, ]
  fit <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident, data = d,
family = "binomial")

  return(coef(fit))
}

# Set seed to replicate results
set.seed(12345)

logit.boot <- boot(data=ccFIT20, statistic=logit.bootstrap, R=10000) # 10'000 samples

logit.boot

plot(logit.boot)


boot.ci(logit.boot, type = "basic", index=1) #intercept
boot.ci(logit.boot, type = "basic", index=2) #FIT result
boot.ci(logit.boot, type = "basic", index=3) # Age
boot.ci(logit.boot, type = "basic", index=4) #Gender

# ------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------
#Performance Measures:

###Compare models by finding the difference in the deviance statistics which is chi-square distributed###

anova(cclog.mod.1, cclog.mod.2, test = "LRT")

#FIT only model
cclog.mod.1 <- glm(CANCERAA~ log(TRANS_FIT_KIT_RESULT+1), ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.1)
exp(cclog.mod.1$coefficients)
exp(confint(cclog.mod.1, level=0.95))

#FIT plus risk model
cclog.mod.2 <- glm(CANCERAA ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident, data
= ccFIT20, family=binomial(link="logit"))

summary(cclog.mod.2)
exp(cclog.mod.2$coefficients)
exp(confint(cclog.mod.2, level=0.95))

# ------------------------------------------------------------------------------------------------------------
#Calibration plot

install.packages("PredictABEL")
library("PredictABEL")
# ------------------------------------------------------------------------------------------------------------
#Calibration Plot (for the risk adjusted model)

cOutcome <- 13 #the column with the outcome in your dataset
predRisk <- predRisk(cclog.mod.2)
# specify range of x-axis and y-axis
rangeaxis <- c(0,1)
# specify number of groups for Hosmer-Lemeshow test
groups <- 10

cal <- plotCalibration(data=ccFIT20, cOutcome=cOutcome, predRisk=predRisk, groups=groups, rangeaxis=rangeaxis)

#Return observed versus expected probability table and p value for Hosmer-Lemeshow goodness of fit test
cal
```

```
#-----------------------------------------------------------------------------------------------------------
#Another way to determine Hosmer Lemeshow statistic
install.packages("ResourceSelection")
library("ResourceSelection")

hl <-hoslem.test(ccFIT20$CANCERAA, fitted(cclog.mod.2), g=10)
hl

#loop to give results for different group splits

for (i in 5:15) {
  print(hoslem.test(ccFIT20$CANCERAA, fitted(cclog.mod.2), g=i)$p.value)
}

# ----------------------------------------------------------------------------------------------------------
#Predictiveness curve

# obtain predicted risks
predRisk1 <- predRisk(cclog.mod.1)
predRisk2 <- predRisk(cclog.mod.2)
# specify range of y-axis
rangeyaxis <- c(0,1)
# specify labels of the predictiveness curves
labels <- c("FIT only", "Risk adjusted")

# produce predictiveness curves
plotPredictivenessCurve(predrisk=cbind(predRisk1,predRisk2),rangeyaxis=rangeyaxis, labels=labels)

# ----------------------------------------------------------------------------------------------------------
# ----------------------------------------------------------------------------------------------------------
#Plotting ROC curve
#Install pROC packages
install.packages("pROC")
library("pROC")

# ----------------------------------------------------------------------------------------------------------
# ----------------------------------------------------------------------------------------------------------
#1st ROC just TRANSFIT

#Predicted probabilities for the two models
#FIT only model
ccFIT20$Predictp1 <- predict(cclog.mod.1, ccFIT20, type = "response")
#Risk adjusted model
ccFIT20$Predictp2 <- predict(cclog.mod.2, ccFIT20, type = "response")

#1st ROC curve for FIT only
roccurve1 <- roc(ccFIT20$CANCERAA ~ ccFIT20$Predictp1)
#Return AUC
roccurve1

#Alternatively for 95% CI:
auc(ccFIT20$CANCERAA, ccFIT20$TRANS_FIT_KIT_RESULT)
ci.auc(ccFIT20$CANCERAA, ccFIT20$TRANS_FIT_KIT_RESULT, conf.level=0.95, method="delong")

#Return AUC CI
ci.auc(roccurve1, conf.level=0.95, method="delong")
ci.auc(roccurve1, conf.level=0.95, method="bootstrap", boot.n = 10000)

# ----------------------------------------------------------------------------------------------------------
#2nd ROC curve for Risk adjusted

roccurve2 <- roc(ccFIT20$CANCERAA ~ ccFIT20$Predictp2)

#Return AUC
roccurve2

#Return AUC CI
ci.auc(roccurve2, conf.level=0.95, method="delong")
ci.auc(roccurve2, conf.level=0.95, method="bootstrap", boot.n = 10000)
# ----------------------------------------------------------------------------------------------------------
```

```
#Combine in a ROC plot

#ROC for risk adjusted model
roccurve2 <- plot.roc(ccFIT20$CANCERAA, ccFIT20$Predictp2)

#Add ROC for FIT only
roccurve1 <- plot.roc(ccFIT20$CANCERAA, ccFIT20$Predictp1, add=TRUE, col="red", lty=3)

#Add legend
legend("right", legend = c("Risk-adjusted LR","FIT only"), lty=c(1, 3), col=c("black","red"))

# ------------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------------
#ROC test - to test for significant difference between ROC curves

roc.test(roccurve1, roccurve2)

# The latter used Delong's test. To use bootstrap test:
roc.test(roccurve1, roccurve2, method="bootstrap", boot.n=10000)

# ------------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------------
#Cross validation functions used for modelling:
#CVfunction.R

# false and true positive rate
ft.pr <- function(mod, thresh = 0.5){
  S <- predict(mod, type = "response")
  Ps <- (S > thresh) * 1
  obs.dat <- as.integer(mod$fitted.values + residuals(mod, type = "response"))
  FP <- sum((Ps == 1) * (obs.dat == 0)) / sum(obs.dat == 0)
  TP <- sum((Ps == 1) * (obs.dat == 1)) / sum(obs.dat == 1)
  vect <- c(FP, TP)
  names(vect) <- c("FPR", "TPR")
  return(vect)
}
# ft.pr(mod = logreg.mod, thresh = 0.5)

# area under curve
auc.roc <- function(mod, resol = 0.001){
  result <- sapply(seq(0, 1, resol), ft.pr, mod = mod)
  span <- dim(result)[2] - 1
  est.auc <- 0
  for(i in 1:span){
    bx.height <- as.numeric((result["TPR", i] + result["TPR", i + 1])/2)
    bx.width <- as.numeric(abs(result["FPR", i] - result["FPR", i + 1]))
    est.auc <- est.auc + bx.width * bx.height
  }
  return(est.auc)
}
# auc.roc(mod = logreg.mod)


# cv false and true positive rate
cvft.pr <- function(obs, pred, thresh = 0.5){
  Ps <- (pred > thresh) * 1
  obs.dat <- obs
  FP <- sum((Ps == 1) * (obs.dat == 0)) / sum(obs.dat == 0)
  TP <- sum((Ps == 1) * (obs.dat == 1)) / sum(obs.dat == 1)
  vect <- c(FP, TP)
  names(vect) <- c("FPR", "TPR")
  return(vect)
}


# cv area under curve
cvauc.roc <- function(obs, pred, resol = 0.001){
  result <- sapply(seq(0, 1, resol), cvft.pr, obs = obs, pred = pred)
  span <- dim(result)[2] - 1
  est.auc <- 0
```

```
  for(i in 1:span){
    bx.height <- as.numeric((result["TPR", i] + result["TPR", i + 1])/2)
    bx.width <- as.numeric(abs(result["FPR", i] - result["FPR", i + 1]))
    est.auc <- est.auc + bx.width * bx.height
  }
  return(est.auc)
}

# cv
cv.logreg <- function(mod, nfold = 10, nreps = 10, seedno = 1234, sig = 4){

  # set-up model
  set.seed(seedno)
  logn <- function(x){if(x == 0){0} else {log(x)}}
  mod.terms <- attr(mod$terms, "term.labels")
  mauc.cv <- mauc <- res.df <- mdev.cv <- mdev <- vector(length = length(mod.terms) + 1)
  resp.loc <- as.numeric(attr(terms(mod), "response"))
  resp.var <- attr(attr(attr(mod$model, "terms"), "dataClasses"), "names")[resp.loc]
  m.dat <- eval(mod$call[["data"]])

  # nreps of nfold cv
  mod.set <- 1:length(mod.terms)
  for(i in 0:length(mod.terms)){
    ii <- i + 1
    pred.cv <- vector(length = dim(m.dat)[1])
    text.form <- paste(paste(resp.var," ~ ", sep = ""), paste(mod.terms[mod.set != i], collapse = " + "), sep = "")
    mod.form <- as.formula(text.form)
    n.mod <- update(mod, formula =  mod.form)
    mdev[ii] <- n.mod$deviance
    res.df[ii] <- n.mod$df.residual
    mauc[ii] <- auc.roc(mod = n.mod)
    for (k in 1:nreps){
      rand <- sample(nfold, dim(m.dat)[1], replace = T)
      for (j in sort(unique(rand))) {
        mod.cv <- update(mod, formula = mod.form, data = m.dat[rand != j,])
        pred.cv[rand == j] <- pred.cv[rand == j] + predict(mod.cv, m.dat[rand == j,], type = "response")
      }
    }
    obs.dat <- as.integer(mod$fitted.values + residuals(mod, type = "response"))
    pred.cv <- pred.cv / nreps
    mdev.cv[ii] <- 2 * sum(obs.dat * sapply(obs.dat/pred.cv, logn) + (1 - obs.dat) * sapply((1 - obs.dat)/(1 - pred.cv), logn))

    mauc.cv[ii] <- cvauc.roc(obs = obs.dat, pred = pred.cv)
  }

  # lrt
  mod.terms <- c("Full", mod.terms)
  lrt <- mdev[2:length(mdev)] - mdev[1]
  lrt.cv <- mdev.cv[2:length(mdev.cv)] - mdev.cv[1]
  df <- res.df[2:length(res.df)] - res.df[1]

  # summary data frames
  mod.sum <- data.frame(mod.terms = mod.terms, res.df = res.df, df = c(NA, df), mdev = round(mdev, sig),
                        lrt = round(c(NA, lrt), sig))
  p.value <- pchisq(lrt, df, lower.tail = FALSE)
  raw.out <- data.frame(mod.sum, p.value = round(c(NA, p.value), sig), mauc = round(mauc, sig))
  cvmod.sum <- data.frame(mod.terms = mod.terms, res.df = res.df, df = c(NA, df), mdev.cv = round(mdev.cv, sig),
                          lrt.cv = round(c(NA, lrt.cv), sig))
  p.value <- pchisq(lrt.cv, df, lower.tail = FALSE)
  cv.out <- data.frame(cvmod.sum, p.value = round(c(NA, p.value), sig), mauc.cv = round(mauc.cv, sig))


 # output
  list(cv = cv.out, raw = raw.out)

}

# cv.logreg(mod = logreg.mod, nreps = 10, seedno = 253636)
```

Appendix 4: Tables of Results for the FIT Participants Adequately Screened (n=27,066)

Latest Event Description (closed episode)	Number of Participants
Subject Discharge Sent (Normal)	24750
Abnormal	528
Low-risk Adenoma	464
Normal (No Abnormalities Found)	266
Intermediate-risk Adenoma	227
GP Discharge Sent (Normal)	200
High-risk Adenoma	197
GP Discharge Sent (No Agreement to Proceed with Diagnostic Tests)	78
Cancer	74
GP discharge letter sent (refusal of positive assessment appointment)	53
Handover into Symptomatic Care - Patient Letter Printed	52
Letter of Non-agreement to Continue with Investigation sent to GP	29
GP Discharge Sent (No show for Positive Appointment)	27
Attended Practitioner clinic	22
1st Positive Appointment Cancellation Requested (Patient to Reschedule)	15
GP Discharge Sent (Unsuitable for Diagnostic Tests)	14
Invited for Diagnostic Test	13
Close Screening Episode via Interrupt	7
Waiting for Clinician Review	7
Decision not to Continue with Diagnostic Test	6
Waiting Decision to Proceed with Diagnostic Test	4
2nd Positive Appointment Cancellation Requested (Patient to Reschedule)	3
Handover into Symptomatic Care, Patient Unfit, GP Letter Printed	3
Suitable for Endoscopic Test	3
1st Positive Appointment Cancellation Requested (Screening Centre)	2
1st Positive Appointment Non-attendance Sent (Patient)	2
DNA Diagnostic Test	2
N/A	2
Not Suitable for Diagnostic Tests	2
Patient Discharge Sent (Unsuitable for Diagnostic Tests)	2
Patient Refused Positive Appointment	2
1st Positive Appointment Booked	1

1st Positive Appointment Cancellation Sent (Patient to Consider)	1
Cancel Diagnostic Test	1
GP Discharge Letter Printed - No Patient Contact	1
GP Discharge letter sent (Discharge by Screening centre)	1
Post-investigation Appointment Invitation Letter Printed	1
Post-investigation Appointment NOT Required - Result Letter Created	1
Post-investigation Contact Made	1
Reminder of Retest Kit Sent (Technical Fail)	1
Suitable for Radiological Test	1
Total – 27,066	

Table A.4.1: Recorded outcomes for participants based on the latest event description field.

Outcome	Numbers for Southern	Numbers for Midland	Total
Abnormal	260	273	533
Cancelled	13	22	35
Cancer	39	35	74
High-risk Adenoma	111	103	214
Intermediate-risk Adenoma	128	135	263
Low-risk Adenoma	226	242	468
Normal (No Abnormalities Found)	138	128	266
Not Attended	121	142	263
Subject Discharge Sent (Normal)	13629	11321	24950
Total	**14665**	**12401**	**27066**

Detection rate for cancer Southern	0.266
Detection rate midlands	0.282
Overall cancer detection	0.273
Overall advanced adenoma + Cancers	2.036
Overall advanced adenoma (HR and IR adenoma)	1.762
Overall all neoplasms (Cancer + HR + IR + LR Adenoma)	3.765
Number attended colonoscopy/diagnostic test. Those with abnormal, cancer, HR, IR, LR, Normal	1818

Table A.4.2: Outcomes for participants reclassified into definitive groups and cancer detection rates.

Appendix 5: Comparison of population with diagnostic follow up versus without.

	Missing diagnostic follow up (n=299) (264 did not attend the appointment 35 people cancelled)	Not missing diagnostic follow up (n=1818)
% Female	39.13	45.27
Mean age	67.33	66.55
Median age	67.00	67.00
Median IMD	17.46	14.64
Mean IMD	21.52	19.31
Median FIT	64.00	55.60
% first time invitee	10.37	10.45
% previous non responders	20.74	13.81
% of previous responder	68.90	75.80

Table A.5.1: Demographics of those missing diagnostic follow up versus those with diagnostic outcomes

Appendix 6: Hosmer-Lemeshow Statistics for Different Group Splits

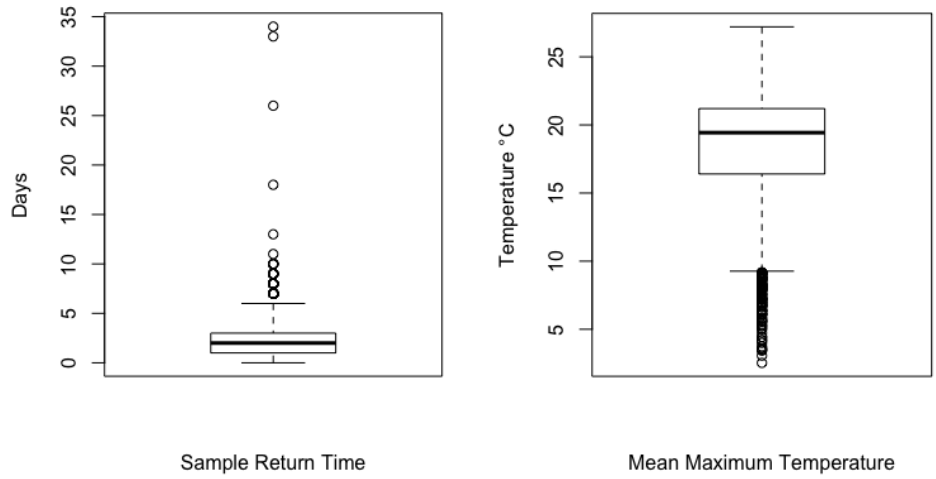
Number of groups	P value for the Hosmer and Lemeshow goodness of fit test
5	0.906
6	0.716
7	0.611
8	0.802
9	0.793
10	0.898
11	0.647
12	0.806
13	0.989
14	0.798
15	0.940

Table A.6.1: Hosmer-Lemeshow goodness of fit test p-values for different group splits for the risk adjusted model

Number of groups	P value for the Hosmer and Lemeshow goodness of fit test
5	0.100
6	0.395
7	0.304
8	0.191
9	0.504
10	0.481
11	0.328
12	0.536
13	0.329
14	0.047
15	0.166

Table A.6.2: Hosmer-Lemeshow goodness of fit test for different group splits for the FIT only model

Appendix 7: Boxplots for Sample Return Time and Mean Maximum Temperature



Sample Return Time

Mean: 2.14, **Median**: 2.00, **SD**: 1.26, **Max:** 34.00, **Min**: 0.00

Mean Maximum Temperature

Mean: 18.92, **Median**: 19.43, **SD**: 3.64, **Max**. 27.20, **Min**. 2.53

Development of a Risk Prediction Model for Colorectal Cancer Screening using an Artificial Neural Network

Chapter based on the following published paper: Cooper, J. A., et al. (2018). "Risk-adjusted colorectal cancer screening using the FIT and routine screening data: development of a risk prediction model." Br J Cancer **118**(2): 285-293.

ABSTRACT

Objectives: Although logistic regression is typically used in medical research for classification and prediction modelling, other statistical methods exist which could provide better model performance and therefore clinical outcomes. The methodology for logistic regression is well developed and the regression coefficients can provide a form of clinical interpretation hence the popularity in employing this method in research. There is however evidence to suggest that other machine learning algorithms such as: decision trees, k-nearest neighbours, artificial neural networks (ANNs) and support vector machines could outperform this conventional method in certain contexts. ANNs tend to have a lower generalization error compared to these other methods and is the next most commonly employed method in the literature. Compared to support vector machines, the ANN provides an absolute risk probability which can be used for individual risk estimates. The real advantage of ANNs over their more standard statistical counterpart are their flexibility and ability to model complex nonlinear relationships between dependent and independent variables. The previous chapter investigated developing a risk prediction model which combined the FIT with routine screening data using logistic regression. The aim of this study was to investigate whether an ANN using the same data improved model performance and test accuracy further when compared to the equivalent logistic regression model.

Design: The same data as the previous chapter (six-month pilot study) were used to train an artificial neural network for risk predictions. This comprised those with a positive FIT (≥20 µg/g) and with a diagnostic colonoscopy outcome (n=1810). A feed forward ANN was developed using a back-propagation algorithm, which attempts to minimize the mean square error for the dataset, with cross-validation. The same risk factors as the previous chapter were investigated for model inclusion (age, sex, Index of Multiple Deprivation score and previous screening history). Network complexity was investigated by assessing weight decay values between 0.01 and 0.1, the number of hidden unit nodes and through

pruning network connections by dropping weights with the lowest magnitude and assessing the change in cross-validated deviance. These methods also improve the generalization of the model in other datasets. Data were normalized before model fitting to improve model performance. Discriminatory power and calibration of the ANN was compared to the logistic regression model using the AUC ROC, Hosmer-Lemeshow statistic and by plotting calibration curves. To compare test accuracy between the logistic regression model and neural network, the ROC curves were plotted and a ROC test was performed to determine if the difference was significant. Sensitivity and specificity were compared at a cutpoint which corresponded to 160 µg Hb/g faeces. Results were also broken down by both outcome severity and sex. Patient profiles are presented for 10 individuals with risk probabilities obtained from the ANN, logistic regression model reported in the previous chapter and the FIT result only.

Results: Standardizing the variables for the model led to lower cross validated deviances. A matrix was produced which assessed different numbers of hidden layer nodes, variables and weight decay parameters through cross validation. A network with 5 input nodes, 3 hidden layer nodes and 1 output node using a weight decay of 0 and 22 weights gave the lowest cross validated deviance and was selected to develop further. A weight decay of 0.01 gave the lowest sum of squared errors and was used in the final model. Network pruning investigations resulted in removing 4 connections to give a lower cross-validated deviance (2077.694) with 18 network weights. The AUC for the Neural Network was 0.69 compared with 0.66 for the logistic regression model. A ROC test confirmed this difference was significantly different ($p<0.001$). Calibration for the ANN assessed using the Hosmer-Lemeshow statistic gave a similar result (0.8924) to the risk adjusted logistic regression model (0.8977) indicating good model fit. At a threshold of 160 µg/g which is the anticipated NHS BCSP cutpoint; the ANN has a sensitivity of 35.15% and a specificity of 85.57% compared to a sensitivity of 33.15% and specificity of 84.69% for the equivalent logistic regression model. Compared with the risk-adjusted logistic regression model, 11 additional advanced adenomas were detected (13 more high risk adenomas, 2 less intermediate adenomas).

Conclusions: Although it is often argued that neural networks are more difficult to interpret and are often likened to a ‘black box’ this study shows the promise of machine learning algorithms for use in screening decisions and clinical practice. Both the logistic

regression and neural network models can give the absolute risk for each individual and this can be used for screening referral decisions by setting an appropriate 'risk threshold'. This neural network uses a logistic activation function and the corresponding risk equation can be provided for external validation and use in screening practice. With the shift to larger and more complex electronic health data, machine-learning algorithms may be better placed to deal with larger amounts of data and non-linear associations when compared with conventional models such as logistic regression.

1.0 INTRODUCTION

The previous chapter developed a risk prediction model which combined the FIT with routine screening data using logistic regression. This model and approach was shown to outperform using the screening test (FIT) on its own in terms of model performance parameters (discrimination and calibration) and test accuracy (sensitivity and specificity at the same number of referrals, at a range of thresholds). Although the standard modelling methodology, logistic regression may not be the optimal statistical methodology in the setting described here. While the resultant model parameters are easy to interpret, this method is limited in many ways, principally by the linearity of the parameters on the scale of the linear predictor (log-odds ratio). An alternative, and possibly better performing model, might be a feed-forward artificial neural network (ANN). This model is highly flexible and, unlike logistic regression, does not require the strong assumption of linearity for combinations of variables and thus allows the inclusion of more complex nonlinear relationships between predictors and the response variable into the prediction model.[1]

1.1 Risk Prediction Models and Machine Learning Algorithms

Clinical prediction rules can be used to assist clinicians in making decisions relating to patient care.[2] This involves the combination of predictors such as patient characteristics and test results in order to estimate the probability of an outcome or the most effective intervention. A model with good performance is one where the probabilities identified from the model match the observed outcomes. Examples of risk prediction models (diagnostic and prognostic) applied in public health and clinical practice include, the Framingham risk equation for cardiovascular disease (used to identify those at high risk of cardiovascular disease)[3] and the Gail Breast Cancer prediction model (used to predict whether a woman will develop breast cancer over a certain time interval)[4]. More recently, risk-scoring systems to automatically identify symptomatic colorectal cancer for referral in primary care have been developed using electronic health records and subsequently tested in a clinical setting.[5 6] Hippisley-Cox and Coupland (2012) have also developed an algorithm to predict the absolute risk of colorectal cancer in primary care, to facilitate referral and early diagnosis.[7]

Typically in medical research, logistic regression is used for classification and predicting binary outcomes. The methodology is well developed and the regression coefficients can

provide a form of clinical interpretation.[8] There are however further methods which can be used for classification purposes including machine learning algorithms such as; decision trees, k-nearest neighbours (white-box models where model parameters can be determined along with logistic regression), neural networks and support vector machines (so called black-box models)[9] which may perform better than conventional methods. Data classification can give just a dichotomous outcome (support vector machine), or an approximation of the probability of an outcome or 'class membership'.

By dichotomizing risk into classes, individual information is lost as probabilities are standardized for all individuals within one group.[10] Whilst support vector machines assign a dichotomous outcome result; logistic regression, ANNs, k-nearest neighbors and decision trees give a probability of class membership but vary in their approximation method.[9] Disadvantages of the k-nearest neighbors algorithm is in defining case neighborhood which is often performed through trial and error, decision trees split continuous variables which is also not recommended as information is lost at each stage. In general, the performance of neural networks and logistic regression models are better than decision trees and *k*-nearest neighbors.[9] Support vector machines have been shown to have similar performance but are not implemented as routinely in statistical software.[11 12] In medicine, neural networks and logistic regression are the most commonly used models which appears to be due to their lower generalization error and ease of model development compared to other models such as support vector machines.[9] This is reflected by the number of publications indexed in Medline for these different types of model: 28,500 for logistic regression, 8500 for neural networks, 1300 indexes for *k*-nearest neighbors, 1100 for decision trees, and 100 for support vector machines.[12]

1.2 Description of Artificial Neural Networks

ANNs are a type of machine learning algorithm which can be considered analogous to the structure and function of biological neurones or the human brain. ANNs consist of a number of interconnected nodes whereby knowledge is obtained through a learning and adaptation process (from a training dataset) and stored as connection weights between nodes.[13 14] Input nodes have the values of predictor variables and are connected to hidden layer nodes through connection weights (which can be considered analogous to the β coefficients in a regression model). Connection weights are most often optimized using a

back-propagation algorithm where the variables for an observation are used to determine predicted output and any difference between predicted output and known output is calculated as an error and back-propagated through the network, this is repeated until overall error is minimized.[1] The hidden layer allows the network to calculate intermediate values and allows non-linear relationships to be modelled between the predictors and outcome. Each node in this layer is then connected to the output node/s. Bias nodes can also connect to the hidden layer or output nodes and can be considered analogous to the intercept parameter in logistic regression. Once the network has been trained it can be used for classification or pattern recognition in a validation dataset. A diagram of a neural network is shown in **Figure 1**.

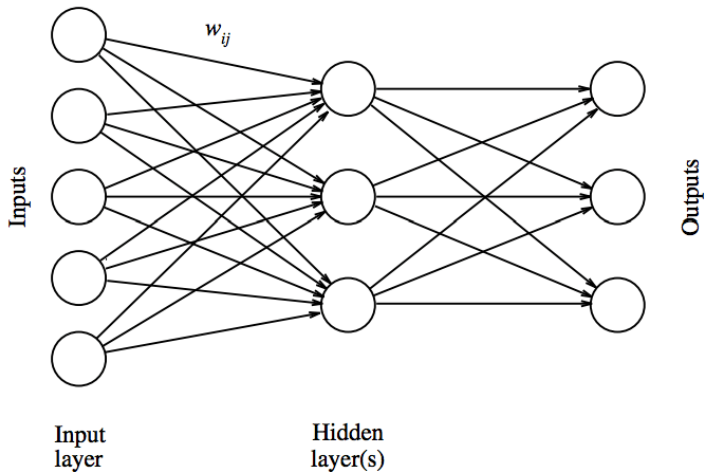


Figure 1: A Feed Forward Neural Network with 5 input nodes (predictors), one hidden layer (with 3 nodes) and 3 output nodes. It is also possible to have connections direct from inputs to outputs (skip-layer connections). The circles are 'nodes' and the lines are connection weights.[15]

1.3 Comparison of ANNs with Logistic Regression

ANNs can be seen as an extension of logistic regression since an ANN with no hidden layer is the same as the equivalent logistic regression model.[1 16] Advantages of ANNs include the ability to model complex nonlinear relationships between dependent and independent variables as well as encompassing all interactions between the variables.[17] Using a logistic regression model for these types of relationships would require further modelling, for example, including an interaction term or taking the log of a continuous variable in order for it to better match linearity. In addition, ANNs require no assumptions about the shape of the underlying probability distribution,[18] meaning risk factors can be included without a

known consideration of their association with the outcome or interactions they may have with other variables.

On the other hand, ANNs are more likely to be over-fitted to the data compared with logistic regression where the model complexity can be considered lower. There are however several approaches that can be taken during model development to limit this including pruning the network, adding a weight decay term and 'early stopping' whereby model fitting is stopped before the maximum likelihood estimate is found.[9]

ANNs are often considered a 'black box' and interpretation is seen as more complex than a standard linear model.[17 19] Logistic regression also offers the advantage of being able to identify possible causal relationships between predictor variable and outcome by providing β coefficients and odds ratios. [1] Furthermore, during model development, variables which are not strongly associated with the outcome can be dropped creating a more parsimonious model. The methodologies for ANN model development are not as formalized, researchers must investigate different training parameters and network architecture to develop the best fitting model.[18]

1.4 Literature Review of Logistic Regression versus ANNs for Medical Datasets

Recent artificial neural networks and other machine learning algorithms have been developed for many clinical areas including cardiovascular prediction,[20 21] breast cancer image recognition and risk calculation,[22 23] lung cancer,[24 25] prostate cancer diagnosis,[26] stroke,[27] psychiatric disorders [28] diabetic retinopathy[29 30] as well as in bioinformatics and laboratory medicine[31 32]. The most recent neural network developed for use in predicting advanced colorectal neoplasia to tailor screening had greater performance than the standard logistic regression model (AUC: 0.721 (95% CI: 0.680-0.762) versus 0.817 (95% CI: 0.789-0.847)) and a more complex deep learning network with additional parameters had even greater performance (AUC: 0.860 (95% CI: 0.837-0.883)).[33] Artificial neural network algorithms have been developed since the 1980s and applied for image recognition in the 1990s.[34] Citations relating to machine learning in health have increased tenfold from 2007-2017 compared to those published up to 2006.[32] It has been suggested that machine learning and artificial intelligence will soon bring a 'paradigm shift' to medical practice.[27 32 35]

A review by Sargent[17] compared the performance of ANNs with regression models for medium-large datasets of medical studies (more than 200 participants). Methods to compare predictive accuracy between the two models was mainly the use of the area under the receiver operating characteristic curve (AUC ROC). In 10 out of 28 studies (36%) the ANN outperformed regression, whereas the ANN was outperformed by regression in 4 cases (14%). In the remaining 14 cases (50%) the methods had similar performances. For the largest 8 studies, the methods tied for 7 cases, with regression having the highest performance for the remaining study. The author suggests that both methods should be explored and used in a complementary manner.[17]

Another review assessed the methodology of 72 papers which compared ANNs with logistic regression for medical datasets.[9] The papers were analysed and rated against several criteria including, dataset size, model parameter selection procedure and performance measures. The model building procedure was reported more often for logistic regression as expected due to the implementation of stepwise methods in software packages. When comparing the discrimination of the models, performance was similar, however the more flexible ANNs outperformed logistic regression in more studies. Lisboa and Taktak[36] conducted a systematic review to determine the clinical benefit of ANNs used for decision support in cancer. The number of clinical trials and randomized controlled trials utilizing ANNs has increased in recent years from 1 to 38.[36] Out of 27 trials identified in the systematic review, 21 showed an increased healthcare benefit. In the remaining 6 studies ANNs performed as well as more traditional statistical methods.

Ahmed[37] carried out a review of ANNs for the diagnosis and survival prediction for colon cancer. The review found that the application of ANNs improved colon cancer classification and survival prediction when compared to more conventional statistical methods but suggest that care must be exercised when reporting or using such models to enhance the confidence in reliability of the data. Other ANNs which have been investigated for colorectal cancer include the use of a model to predict distant metastasis. Comparing the LR model to the ANN in this setting revealed that the accuracy was greater for the ANN (measured through the AUC ROC), 0.81 versus 0.78.[38] The data for this study were however split into a train and test dataset for the ANN which is thought to be an inefficient form of internal validation as it does not use all the available data to build the model.[39]

The systematic review reported in **Chapter 2** identified that machine learning algorithms have not been previously developed which combine the FIT with other risk factors for colorectal cancer screening. Other systematic reviews of risk prediction models for colorectal cancer have been carried out for both asymptomatic and symptomatic individuals.[40 41] A recent systematic review of risk prediction models which allow the identification of people at higher risk of colorectal cancer with symptoms was performed by Williams *et al*. (2016).[41] For study selection, the risk prediction model had to include one or more individual risk factors (including symptoms) for undiagnosed colorectal cancer. The review identified 18 papers (15 models) for inclusion; 9 for primary care and 6 for secondary care. Four of these studies required questionnaires or interviews to obtain data on risk factors and 11 needed the input of a health-care worker. Four models included the guaiac based FOBT within the model (BB Equation and CAPER score).[5 42-44] None of these models were designed specifically to be used as part of a screening programme but were instead developed for primary care case finding for urgent referral. Furthermore, the models which included FOBT all used logistic regression for model development. Clinical utility of a risk model was only assessed by one study using a risk score developed by Hamilton et al.[45]

A review of risk prediction models for identifying the future risk of colorectal cancer in asymptomatic individuals was carried out by Usher-Smith *et al*.[46] Forty papers describing 52 risk models and 6 external validation studies were included for analysis. Six of the risk models included variables only available in routine medical records whereas the majority (n=32) utilized a questionnaire for data. Of the models using routine data for advanced colorectal neoplasia, the best performing logistic regression model was used for selecting individuals for a primary screening test of colonoscopy and included age, gender and BMI (AUC ROC, 0.65).[40] No models identified in this review used machine learning algorithms.

1.5 Potential Barriers to the Implementation of Neural Networks and Other Machine Learning Algorithms to Healthcare

Although there have been many studies which have developed machine learning algorithms for a clinical application, few have been implemented in practice. This is in spite of the potential for increased diagnostic accuracy and model performance and the fewer data assumptions required for these approaches. Medical statisticians are often wary

about the application of machine learning approaches.[47] Medicine also lags behind other industries, which have implemented machine learning approaches widely.[48 49] Several recent articles have discussed potential obstacles to the wider adoption and clinical application of these approaches in healthcare.[32 50-53]

Some of the reasons discussed by Cabitza *et al.*[51] include deskilling clinicians and other healthcare workers by reliance on automated software, intrinsic uncertainty of medical data, the clinical context not being represented by the models and the black box nature of these algorithms.[51] An example of deskilling provided was computer-aided detection for breast cancer screening whereby the sensitivity of readers decreased by 14% when CAD prompts were provided.[54] The authors conclude that before the adoption of machine learning algorithms, evidence is required for improved patient outcomes and acceptability to patients and clinicians along with the performance metrics.[51]

Several responses have been published to this article discussing that the problems of using machine learning in practice are similar to those found in the introduction of other new technologies.[50] Other responses argue that overreliance on automated technology would be due to lack of methodological transparency and that research should focus on making these methods more accessible to healthcare users.[55] Lasko *et al.*[56] infer that negative consequences of applying machine learning approaches can be due to the misapplication of these methods in healthcare and suggest the need for trained medical data scientists with an understanding of both medicine and computer science.

A frequently cited reason for machine learning models not being implemented is their 'black box' nature. Users of the models are unable to interpret outputs for clinical meaning.[51 57] Although models such as neural networks can provide greater accuracy for certain clinical scenarios, clinical interpretability of models could be improved using different methods such as visualisation tools (e.g. Garson's algorithm to show relative importance of variables).[51] The addition of clinicians trained in data science who can contribute to model development has also been suggested as a solution to this current issue. [35 56]

A further reason for the slower progression of machine learning into practice over more standard statistical models is the nature of the data used for these algorithms. Machine learning not only uses standard clinical data but can also use data from images, genetic data or even employ natural language processing.[27] These algorithms are also often developed using large electronic health records. The data used for these models can therefore be much larger, more complex and diverse, making the algorithms more difficult to validate (where the same predictors are required) and to employ as a holistic approach in computer software (e.g. for a diagnosis).

Machine learning algorithms have also been described as 'data hungry' - which refers to predictive performance in relation to sample size.[35 58] Simulation studies comparing the performance of support vector machines and neural networks with standard approaches (logistic regression) found that these models may need 10 times as many events per variable for smaller optimism and stable discrimination measures.[59] In order to perform optimally, machine learning systems also need to be trained regularly which requires a continuous supply of data for model improvement.[27] The complexity of data from diverse sources and the difficulty in implementing or validating another researcher's algorithm has therefore hindered progress in the implementation of these approaches in practice.

Frank E Harrell Jr. has written a number of recent articles on the topic of machine learning and the potential issues in applying these models in medical research.[60] For example, Harrell discusses that researchers often employ machine learning in large electronic health records whereby bias in the healthcare system can be 'perpetuated' for future medical decisions.[60] Another obstacle against implementation is a lack of regulatory standards to assess the effectiveness and safety of 'artificial intelligence systems'. This has been improved in the US recently with the FDA providing guidance to evaluate these systems.[27]

A series of recent articles discuss the theory and application of probability estimation using machine learning.[47 57 58 61-63] Issues of using machine learning approaches over regression models include the need for updating the model in a new clinical context (e.g. secondary versus primary care) or in a different population with varying case mix (e.g. geographical differences). Updating the model for regression involves adjusting the model intercept and is commonly performed. However, it is uncertain and unknown how recalibration of the

model would be performed for a machine learning algorithm,[58 62] a feature which could be deemed essential for validation and applying in clinical practice.

Kruppa *et al.*[63] and Steyerberg *et al.*[58] also discuss the frequently cited issue of interpretability of the model to a clinical audience and whether a clinician will trust a 'noninterpretable machine'. Logistic regression provides interpretable parameter estimates and odds ratios can be reported to assess association with a particular outcome. Variable Importance Measures (VIMs) can be provided for Random Forest Models and allow variables to be ranked by their importance but the magnitude of the estimates have no underlying meaning.[63] Garson's algorithm for neural networks also ranks relative importance but provides no further clinical interpretation.[64] Further research into methods for interpretability need to be undertaken to improve acceptance of these approaches for use in practice. Increasing use of web-based or computer based risk calculators can facilitate the use of these more complex models.[58]

Machine learning approaches can have numerous benefits including statistical consistency of outcome probabilities and extensive flexibility.[63] However, model tuning is also more complex with more tuning options adding an additional layer of complexity and making comparison of different models problematic.[62] The benefits of the flexibility of these models also comes with the problem of a lack of standardization for model development and for use in practice. The methodology for model tuning is not fully developed for artificial neural networks. More modern methods have been developed for logistic regression models to provide additional flexibility including multivariable fractional polynomials, restricted cubic splines and shrinkage approaches.[58 62] Steyerberg *et al.*[58] suggest that machine learning will have a supplementary role but will not replace standard statistical methodology.

Computational transportability is another issue of applying machine learning models in practice.[47] This involves the 'transport and exchange of computer programs allowing prediction' between the model developers and for those who apply or validate the model.[47] Logistic regression for instance is easy to 'transport' since the probabilities are straightforward to obtain with minimal computer software requirements and beta coefficient estimates along with the intercept can be provided for reporting the model. Machine learning algorithms however can often only be applied to their own data if

researchers have the corresponding software objects particularly with models such as Random Forests.[47] Changes in the software could also affect the application of these machine learning models; applying the same models in different software can also yield different results complicating external validation.[47]

Along the pathway of assessing whether a model is fit for practice, a model development study should be followed by external validation studies in either temporal or geographically different datasets and ideally by other research teams to reduce bias. Once externally validated, a model impact study can assess the effect on patient outcomes and provide confidence in the model results and performance before being implemented in practice.[65-68] There are many model development studies, fewer external validation studies and very few model impact studies.[10] External validation of machine learning algorithms by other researchers is much scarcer than those assessing standard statistical techniques. Along with computational transportability as discussed above, this is most likely due to the underlying model being more complex and therefore the model equation is less likely to be provided in published model development studies. TRIPOD guidelines focus mainly upon regression techniques, additional guidance would be required for the unique issues encountered by machine learning techniques.

1.6 Rationale

As far as can be identified there are no studies which have developed an ANN for use in screening referral decisions for colorectal cancer. The studies identified above which have included a screening test result were either developed for case finding in primary care, employed the dichotomous gFOBT or used logistic regression to develop the risk prediction model. There is some evidence to suggest ANNs outperform logistic regression in certain scenarios within medical studies and it has been suggested that both methods should be employed in a complementary manner. Other machine learning methods could be considered for developing a risk prediction model within the context of colorectal cancer screening referrals. However, ANNs and logistic regression are the methods most commonly employed in the medical literature. Drawbacks of other methods such as decision trees include the splitting of continuous variables which results in the loss of information at each stage. The performance of ANNs and logistic regression models have been shown to be generally greater than decision trees and *k*-nearest neighbors. Support vector machines on the other hand assign a dichotomous outcome which results again in

the loss of information in terms of absolute risk prediction and individualized probability for a patient. The real advantage of ANNs are in their flexibility and ability to model complex nonlinear relationships between dependent and independent variables.

Reporting and methodology is less well defined for ANNs and so care must be exercised when developing an ANN so performance measures can be understood fully. The main criticism is that they are considered a 'black box' and that interpretation of network weights is much more difficult than interpreting the coefficient estimates in a logistic regression model. However, weight saliency (the relative importance of weights) can be used, not only for optimizing network structure and performance, but also for giving some intuition about which inputs contribute most to the accuracy of the network predictions.

The aim of this study was to investigate whether developing an ANN which combines FIT with other routine screening data available on the BCSS improves model performance and test accuracy further for an average risk English screening population when compared to the equivalent logistic regression model.

2.0 METHODS

A similar approach to model development was taken to the logistic regression model to aid comparability. The TRIPOD guidelines and STARD statement were used when reporting this study and to form the data analysis plan.[10 69]

A multi-layer perception model which has an input layer, a hidden layer and an output layer was compared to the logistic regression model that was developed in the previous chapter (which combined age, sex, previous screening history and the FIT result). Model fitting proceeded in a similar fashion to that described previously for the logistic regression model using cross-validation, allowing performance to be compared directly.

2.1 Study population and data source

This study used the same data as the model development study discussed in the previous chapter. Briefly, these data were collected for the NHS BCSP comparative study which compared the acceptability and accuracy of the FIT compared to the gFOBT.[70] Data for the FIT only were used where 40,930 individuals were invited to complete a test (one out of

every 28 screening invitations) and 27,167 were adequately screened. This analysis used complete cases and those who had a FIT result of 20 µg Hb/g faeces and above (n=1810). Twenty µg Hb/g faeces was chosen as the cutoff for test positivity during the pilot study and therefore these participants would have been referred on for colonoscopy for a definitive diagnosis.

2.2 Routinely Available Predictors and Test Results

The candidate predictors investigated to be included in the model were the same as the predictors investigated in the previous study. These included age, sex, IMD score and previous screening history (i.e. whether someone was a previous non-responder/responder to screening). These variables were coded in the same way for the model and used the equivalent definitions. The screening test investigated was the OC-SENSOR FIT (Eiken Chemical Co. Ltd., Japan, supplied by Mast Diagnostics, UK) along with the OC-SENSOR Diana analyser. The FIT units were converted from ng Hb/ml buffer to µg Hb/g faeces as recommended by the World Endoscopy Organisation.[71] Those participants with a positive result were referred on for colonoscopy assessment within 14 days of an appointment with a specialist screening practitioner. Alternative investigations were arranged if the colonoscopy was inappropriate for a patient or if the test failed first time round e.g. CT scan or flexible sigmoidoscopy.[72]

2.3 Model Outcome

The model outcome was colorectal cancer or advanced adenoma detected at colonoscopy after a positive FIT referral. Advanced adenomas were those classified as either high-risk or intermediate risk, since these have potential to develop into bowel cancer if untreated, particularly as age increases. [73 74]

2.4 Statistical Analysis

The FIT pilot data provided by the HSCIC through the Office for Data Release (ODR) were analysed in RStudio Version 0.99.903 (driven by R version 3.3.1) on a Windows 7 computer.[75] Additional packages were also loaded from The Comprehensive R Archive Network (CRAN; https://cran.r-project.org/).[76-82] For neural network development, the package 'nnet' developed by Ripley[79] was loaded and used for analysis purposes. Although this package is limited to one hidden layer there is evidence to suggest that additional

layers do not significantly improve the performance of the model,[15] [83] and that most functions encountered in medicine can be modelled using a Perceptron with one hidden layer.[18] Furthermore, the more variables and the more complex the model, the greater the computational time which with large datasets can also lead to convergence problems. 'Neuralnettools' was a package used for neural network visualization and for further analysis.[80] The R scripts used to develop the neural network and assess the performance of the models is provided in **Appendix 1.**

A feed forward ANN was developed using a back-propagation algorithm. Back propagation is a learning mechanism which attempts to minimize the mean square error for the dataset.[18] This is achieved by comparing the output of the network based on the input variables with the true output and then this error is propagated backwards through the network adjusting connection weights appropriately to reduce mean square error.[18] There are multiple training algorithms which could be used for neural network development, however most medical studies have utilized this particular algorithm.[1] A logistic activation function acts on the weights connecting the input nodes (in our case the predictor variables) to the hidden layers, which are then linked to the outputs by an output activation function, which for this binary outcome of colorectal cancer/advanced adenoma status is also logistic. Different activation functions also exist including linear functions and hyperbolic tangent functions; if the output is continuous a linear function for the output node can be selected.[1] The output ranges from 0 to 1 based on the network prediction of the outcome. The non-linearity added by the activation function allows much more general and flexible relationships to be formulated between inputs and outputs. The weights in the ANN model are analogous to parameters in the logistic regression model, but because the relationship between inputs and outputs is more complex they cannot be interpreted in isolation in the same simple manner as in a logistic regression model. A multi-layer ANN model with an input layer (consisting of the same predictors as the logistic regression model), a single hidden layer and an output layer with a single node was compared to the logistic regression model.

2.4.1 Model Development

Model fitting proceeded in a similar fashion to that described for the logistic regression model using cross-validation, allowing performance to be compared directly.[84] The same risk factors (age, sex, IMD score and previous screening history) investigated for model inclusion were included in developing the neural network regardless of whether they were found to be significant in the logistic regression model.

Data Normalisation

The data were normalized before model fitting as this has been shown to improve the performance of the network.[1] Scaling the data in this way, prevents premature saturation of hidden nodes and stops larger numbers overriding smaller ones.[83] The continuous variables were standardised using Gaussian normalization (subtracting the mean and dividing by the standard deviation) and compared to models which did not use standardisation. By normalising the weights, the weight decay parameter became less important and influential in the model.

Network Architecture

Unlike logistic regression where the methods for model development are fully developed, neural networks are not routinely used in medicine and so methods are by necessity more individualised for each study and require more experimentation.[18] Network complexity can be manipulated by changing the weight decay parameter, changing the number of hidden unit nodes and cutting out network connections/links (pruning).[15] A matrix was produced which determined the ten fold cross validated deviance, at varying weight decays, routine predictors and number of hidden nodes (the more hidden layer nodes the more complex the model) in order to determine the optimal (most parsimonious) neural network model architecture.

Neural networks can also be over-trained (over-fitted) to the data, therefore a penalty term can be applied during model optimisation $E + \lambda C(f)$, where E is the fit criterion to minimize during network model fitting, 'C' is a penalty on the roughness of the continuous function 'f' and 'λ' is the weight decay.[14 15] In addition, the maximum number of iterations can be set at an appropriate level and the network 'pruned' to drop any unnecessary

connections. Reducing complexity in this way enables improved generalisation of the model which would be needed when applying it in external populations.[85] The optimum model determined from the cross validated deviance was further refined by investigating different weight decay values and by pruning weight connections based on absolute magnitude. The maximum number of iterations was selected which allowed the models to converge (500).

Ripley suggests the use of weight decay (λ) values between 0.01 – 0.1 for the entropy fit,[14] higher values of weight decay do however improve the stability of the model. Weight decay helps to smooth the model output through regularization, limiting the magnitude of the weights and can be considered analogous to shrinkage in logistic regression.[9] Models without weight decays are more sensitive to changes when weights are dropped; the weight decay level between 0 and 1 was investigated to give the lowest SSE (sum of squared errors/model residuals).

The model was 'pruned' by dropping out weights with the lowest magnitude and assessing the change in cross-validated deviance. Pruning is described by Ripley as a method 'used for removing parts of trees and networks with the aim of increasing generalization'[15]. By removing weight connections within the neural network, the degrees of freedom decrease increasing generalization. A plot of cross-validated deviance against the number of weight connections was produced to assess this.

2.4.2 Model Performance

Discrimination and calibration of the ANN were compared to the logistic regression model using the AUC ROC, Hosmer-Lemeshow statistic and by plotting calibration curves.[86] Discrimination assesses the ability of the model to distinguish between those at high risk of colorectal cancer/advanced adenoma versus those at low risk.

The Hosmer-Lemeshow test determines if there are significant differences between observed and expected numbers using a chi squared test.[87] For this test, observations were split into a different number of risk based groups (between 5 to 15) as this can affect the corresponding result of the Hosmer-Lemeshow statistic.[88]

2.4.3 Test Accuracy of the Risk Model

To assess test accuracy, the ROC curves of ANN and risk-adjusted logistic regression models were compared. As a baseline comparator, the FIT only model (equivalent to just using FIT as a test) was also displayed on the ROC curve. A ROC test to compare the AUC ROC between both the models was performed using Delong's method and bootstrapping. Delong's method compares the AUCs of two or more correlated ROC curves which are constructed from tests/models performed on the same individuals and takes into account the correlated nature of the data.[89]

To assess the sensitivity and specificity of the ANN at specific thresholds, 2 by 2 tables were produced for a threshold of 160µg Hb/g faeces (and the equivalent risk threshold) as investigated for the logistic regression model in the previous chapter. The NHS Bowel Cancer Screening Programme plans to adopt a threshold of 160 µg Hb/g faeces and Wales 150 µg Hb/g faeces. Results are also presented for internationally used lower thresholds 30-50 µg/g [90] and also for 80 µg/g which is the cutpoint which Scotland plans to adopt (Stephen Halloran, personal communication).

Results were broken down by both outcome severity (CRC, high-risk adenoma, intermediate risk adenoma, low-risk adenoma, abnormal and normal (no abnormalities found)) and sex (male and female) for the ANN and logistic regression models using the FIT only as baseline for comparability.

2.4.4 Clinical Utility

The predictiveness curve has been proposed by Pepe *et al.,*[91] as an alternative plot which allows both the assessment of the fit of the model and the clinical utility when applied to the population. It is argued that the predictiveness curve can give additional information about risk-threshold which is not typically provided by the ROC curve. The predictiveness curve was plotted for the ANN and LR model to aid comparison.

Patient profiles for 10 individuals are presented with the corresponding probabilities of cancer/advanced adenoma being detected at colonoscopy estimated from the ANN, logistic regression (LR) model and FIT result only. The profiles allow a more detailed investigation into which variables may push an individual over a referral threshold. To

complement this, a plot using Garson's algorithm to show the relative importance of each variable is presented for the ANN.[64] This approach enables interpretation of neural network connection weights which is seen as a drawback of ANNs when compared to odds ratios for logistic regression.

3.0 RESULTS

3.1 Study Population

Participants with a FIT result of ≥20 µg/g and a definitive diagnostic outcome were used to develop the ANN (n=1810; 549 cases, 1261 with a negative/low risk outcome). The mean age was 66.54 years and 45.3% of participants were females. See **Chapter 3** for more detail on the study population.

3.2 ANN Model Development

3.2.1 Standardisation

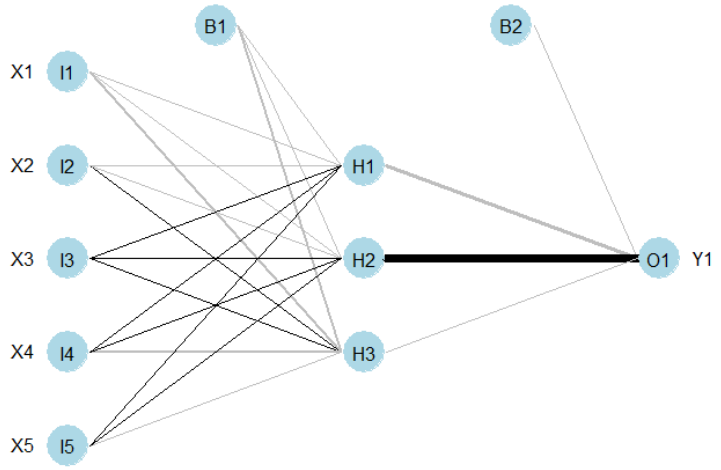
The continuous variables were standardised using Gaussian normalization and compared to models which did not use standardisation. By having no hidden layer, the ANN is equivalent to the logistic regression model as seen in the cross-validated deviances which are the same for a standardized neural network and logistic regression model (**Table 1**). The neural network with standardised variables was more efficient and produced lower cross-validated deviances (signifying better fit) when compared to the model with unstandardized continuous variables as found in previous research (see **Table 1**).

Model Type	Standardised continuous variables?	Deviance	Cross-validated Deviance
Neural Network (5-0-1)	No	Inf	2308.924
Neural Network (5-0-1)	Yes	2103.021	2114.95
Logistic Regression risk-adjusted model	No	2103.000	2113.995
Logistic Regression risk-adjusted model	Yes	2103.000	2113.995
Logistic Regression FIT only model	No	2153.6	2157.825

Table 1: Comparing a standardized ANN (5-0-1) with a non-standardized ANN and the risk-adjusted logistic regression model. The 5-0-1 ANN has 5 input nodes, no hidden layer and one output node with a logistic output function. An infinite value is seen with the neural network as the starting seed/point used may not allow the model to converge within the number of iterations set for the model

3.2.2 Number of hidden nodes

A matrix was produced which assessed different numbers of hidden layer nodes, variables and weight decay parameters and the effects on the cross validated deviance. A network with 5 input nodes, 3 hidden layer nodes and 1 output node using a weight decay of 0 and 22 weights gave the lowest cross validated deviance and was selected to develop further. The 5 input nodes included the same variables as in the final logistic regression model; FIT, age, sex, previous screening history (as two nodes since this variable is treated as a factor). The architecture of this neural network is shown below in **Figure 2**, the weight connection values for each node is included in **Appendix 2**.



Node	Label
I1	Input Node 1 – Standardised FIT result (continuous)
I2	Input Node 2 – Standardised age (continuous)
I3	Input Node 3 – Sex (Factor Male compared to Female)
I4	Input Node 4 – Previous non responder compared with a first time screen (Factor)
I5	Input Node 5 – Previous responder compared with a first time screen (Factor)
H1	Hidden Layer Node 1
H2	Hidden Layer Node 2
H3	Hidden Layer Node 3
B1	Bias Node 1
B2	Bias Node 2

Cross Validated Deviance *(2103.04)* ***Deviance*** *(2048.25)* ***E*** *(1024.12),* ***Penalty*** *(0)****, E.crit*** *(1024.12)*

Figure 2: Architecture of the feed forward 5-3-1 neural network with 22 weights, 500 iterations and 0 weight decay. Neural network plotted using nnet and the neuralnetworktools packages in R. Positive connection weights are represented with black lines, negative connections are represented with grey lines.

3.2.3 Weight Decay

Different values of the weight decay parameter affect the sum of squared errors (SSE), mean squared error (MSE) and the cross-validated deviance. **Table 2** reveals a weight decay of 0.01 has the smallest SSE (346.0445) when investigating weight decay values between 0.0 to 0.5. The smallest cross-validated deviance from this range is obtained using a weight decay of 0.001 (2100.892).

Weight Decay	SSE	MSE	Cross Validated Deviance	E	Penalty	E.Crit
0.0001	347.060	0.192	2103.719	1025.033	0.833	1025.866
0.001	348.023	0.192	2100.892	1027.890	1.684	1029.574
0.01	346.038	0.191	2103.489	1023.470	3.175	1026.646
0.1	350.959	0.194	2107.905	1035.979	4.078	1040.056
0.2	351.486	0.194	2106.162	1037.336	6.060	1043.396
0.3	352.005	0.194	2104.963	1038.673	7.474	1046.147
0.4	352.663	0.195	2104.912	1040.304	8.062	1048.367
0.5	353.135	0.195	2104.899	1041.407	8.847	1050.254

Table 2: Changes in weight decay on SSE, MSE, cross validated deviance for a 5-3-1 neural network model.

Investigating values of the weight decay parameter between 0.0 and 1.0 (in 0.0001 increments) the lowest SSE (346.06) is achieved from a weight decay of 0.0102 (**Figure 3**). Looking at a more restricted range of weight decay values (0.0001 and 0.1), a weight decay of 0.0092 gives the lowest SSE (345.97). Based on these investigations, a weight decay of 0.01 was used for the final model.

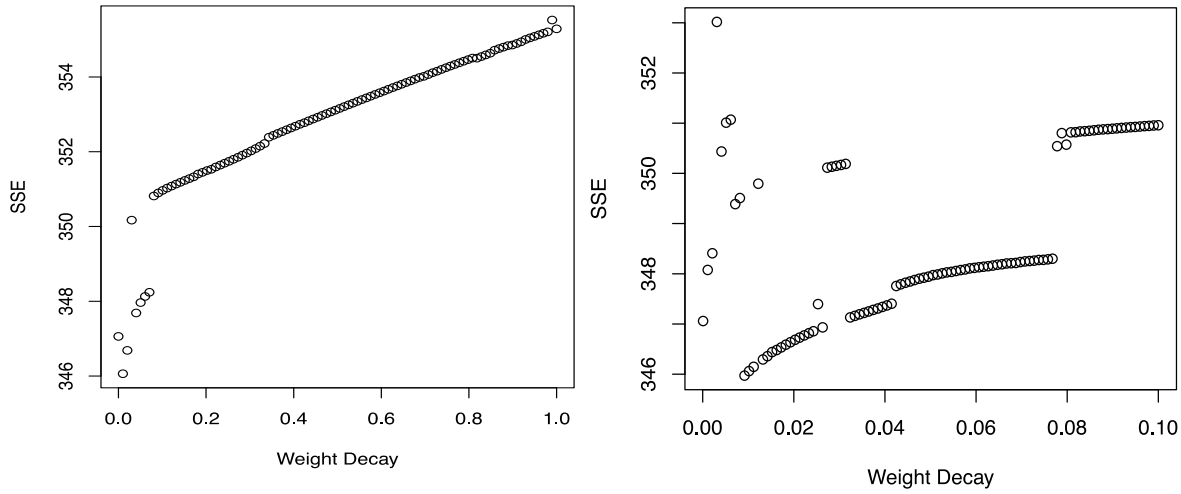


Figure 3: Investigating the effect of changing the weight decay parameter value on the sum of squared errors of the neural network model for weight decay values between 0.0-1.0 (left figure) and then from a more restricted range between 0.0001 and 0.1 (right figure).

3.2.4 Network Pruning

Removing different combinations of weight connections was investigated, small magnitudes were removed and the effect on the resultant cross-validated deviance analysed (see **Figure 4** and **Table 3**). The model with the lowest cross validated deviance (2077.694) was one which removed the following connections; i3->h1, i1->h2, i3->h3, i4->h3.

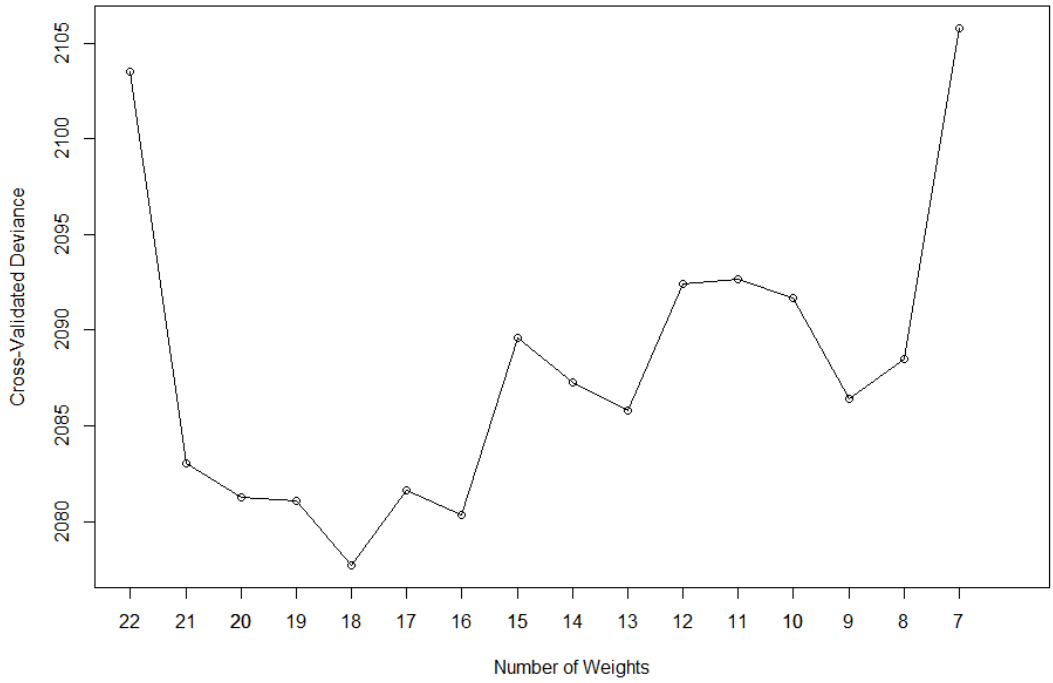


Figure 4: Change in cross-validated deviance as weight connections are dropped from the neural network model.

Number of Weights	Cross Validated Deviance	Weight connection removed	Deviance	E	Penalty	E.crit
22	2103.489	Full model	2046.940	1023.470	3.175	1026.646
21	2083.051	Connection 8	2046.991	1023.496	3.154	1026.649
20	2081.277	Connection 4,8	2047.205	1023.603	3.267	1026.869
19	2081.100	Connection 16,4,8	2048.226	1024.113	3.051	1027.164
18	2077.694	Connection 17, 16, 4, 8	2048.520	1024.260	2.925	1027.185
17	2081.610	Connection 19,17,16,4,8	2052.037	1026.019	2.355	1028.374
16	2080.341	Connection 13,19,17,16,4,8	2052.881	1026.441	2.458	1028.899
15	2089.588	Connection 2, 13,19,17,16,4,8	2059.705	1029.852	2.118	1031.970
14	2087.297	Connection 15, 2, 13,19,17,16,4,8	2063.545	1031.772	2.561	1034.333
13	2085.811	Connection 18, 15, 2, 13,19,17,16,4,8	2063.864	1031.932	2.642	1034.574
12	2092.438	Connection 10, 18, 15, 2, 13,19,17,16,4,8	2076.493	1038.247	3.879	1042.126
11	2092.681	Connection 6, 10, 18, 15, 2, 13,19,17,16,4,8	2076.603	1038.302	3.853	1042.155
10	2091.677	Connection 1, 6, 10, 18, 15, 2, 13,19,17,16,4,8	2076.158	1038.079	4.104	1042.184
9	2086.436	Connection 12, 1, 6, 10, 18, 15, 2, 13,19,17,16,4,8	2076.173	1038.086	4.098	1042.184
8	2088.516	Connection 5, 12, 1, 6, 10, 18, 15, 2, 13,19,17,16,4,8	2079.737	1039.869	3.884	1043.752
7	2105.806	Connection 20, 5, 12, 1, 6, 10, 18, 15, 2, 13,19,17,16,4,8	2094.461	1047.231	4.431	1051.661

Table 3: The order and effect of removing different weight connection values from the ANN on cross validated deviance.

3.2.5 Refined ANN Final Model

The final model which is compared to the risk adjusted logistic regression model is presented in **Figure 5** below (and **Table 4** for weight connection values). This feed forward ANN has 5 input nodes, 3 hidden layer nodes and 1 output node along with two bias nodes providing the added flexibility of the model. The weight decay parameter value was 0.01 and after pruning, the number of network weights was 18. This final refined model was used to assess model performance and test accuracy. The full model equation is given below in **Equation 1.**

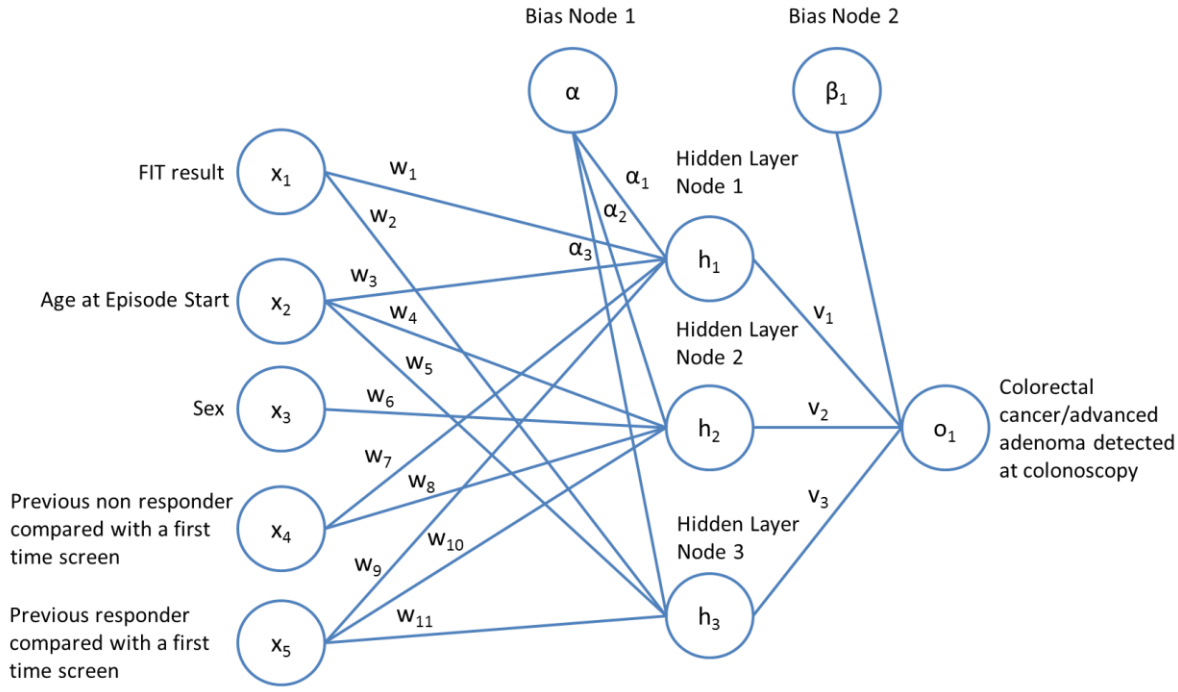


Figure 5: Feed forward 5-3-1 neural network with 18 weights and a weight decay of 0.01. The log of the FIT result and age were normalised before modelling with the neural network.

Node	Formula	Weight Connection Value
b->h1	α1	-3.69
i1->h1	x1w1	-0.23
i2->h1	x2w3	-2.12
i3->h1	/	0.00
i4->h1	x4w7	3.42
i5->h1	x5w9	3.50
b->h2	α2	-3.39
i1->h2	/	0.00
i2->h2	x2w4	-0.87
i3->h2	x3w6	0.35
i4->h2	x4w8	2.11
i5->h2	x5w10	1.72
b->h3	α3	1.06
i1->h3	x1w2	-5.62
i2->h3	x2w5	3.30
i3->h3	/	0.00
i4->h3	/	0.00
i5->h3	x5w11	-5.72
b->o	β1	-0.60
h1->o	v1	-5.11
h2->o	v2	11.25
h3->o	v3	-1.07

Table 4: Weight connection values for the final 5-3-1 neural network model with 18 weights and a weight decay of 0.01.

Probability of colorectal cancer/advanced adenomas being detected at colonoscopy:

$$p = \frac{e^{o_1}}{1 + e^{o_1}}$$

Where:

$$o_1 = v_1 h_1 + v_2 h_2 + v_3 h_3 + \beta_1$$

$$h_1 = \frac{e^{\alpha_1 + \sum w_i x_i}}{1 + e^{\alpha_1 + \sum w_i x_i}}$$

$$h_2 = \frac{e^{\alpha_2 + \sum w_i x_i}}{1 + e^{\alpha_2 + \sum w_i x_i}}$$

$$h_3 = \frac{e^{\alpha_3 + \sum w_i x_i}}{1 + e^{\alpha_3 + \sum w_i x_i}}$$

Applying the final neural network model:

$$p = \frac{e^{o_1}}{1 + e^{o_1}}$$

Where:

$$o_1 = (-5.11)h_1 + (11.25)h_2 + (-1.07)h_3 + (-0.60)$$

$$h_1 = \frac{e^{-3.69+(-0.23)x_1+(-2.12)x_2+3.42x_4+3.50x_5}}{1 + e^{-3.69+(-0.23)x_1+(-2.12)x_2+3.42x_4+3.50x_5}}$$

$$h_2 = \frac{e^{-3.39+(-0.87)x_2+0.35x_3+2.11x_4+1.72x_5}}{1 + e^{-3.39+(-0.87)x_2+0.35x_3+2.11x_4+1.72x_5}}$$

$$h_3 = \frac{e^{1.06+(-5.62)x_1+3.30x_2+(-5.72)x_5}}{1 + e^{1.06+(-5.62)x_1+3.30x_2+(-5.72)x_5}}$$

p = Probability; o_1 = Output 1; a_1 = Bias node 1 to hidden layer 1 weight value; a_2 = Bias node 1 to hidden layer 2 weight value; a_3 = Bias node 1 to hidden layer 3 weight value; β_1 = Bias node 2 to output weight value; x_1 = Standardised log(FIT Result +1); x_2 = Standardised Age at episode start; x_3 = Sex (male compared to female at baseline); x_4 = Previous non responder (compared to first time screen); x_5 = Previous responder (compared to first time screen); h_1 = Hidden layer node 1 intermediate output; h_2 = Hidden layer node 2 intermediate output; h_3 = Hidden layer node 3 intermediate output; v_1 = Hidden layer 1 to output 1 weight value; v_2 = Hidden layer 2 to output 1 weight value; v_3 = Hidden layer 3 to output 1 weight value.

Equation 1: Equation for the risk scores/probabilities obtained from the final 5-3-1 neural network with 18 weights and a weight decay of 0.01. The log of the FIT result and age were normalised before modelling with the neural network. When applying the neural network in new data, the FIT result and age need to be standardised using the standard deviation and mean parameters defined in the training dataset (Age; SD =4.22 , Mean= 66.54 Log of FIT; SD = 1.10 , Mean = 4.35).

3.3 Model Performance

Discrimination

The AUC for the ANN was 0.69 (0.66-0.71) compared with 0.66 (95% CI: 0.63-0.69) for the logistic regression model. A ROC test using bootstrapping with 10,000 iterations shows that the AUC is statistically significantly different (D = -3.5057, p-value = 0.0005). This was validated using Delong's method[89] (Z = -3.5134, p-value = 0.0004). A ROC curve of the ANN compared to the risk adjusted logistic regression model with FIT only at baseline is given in **Figure 6.**

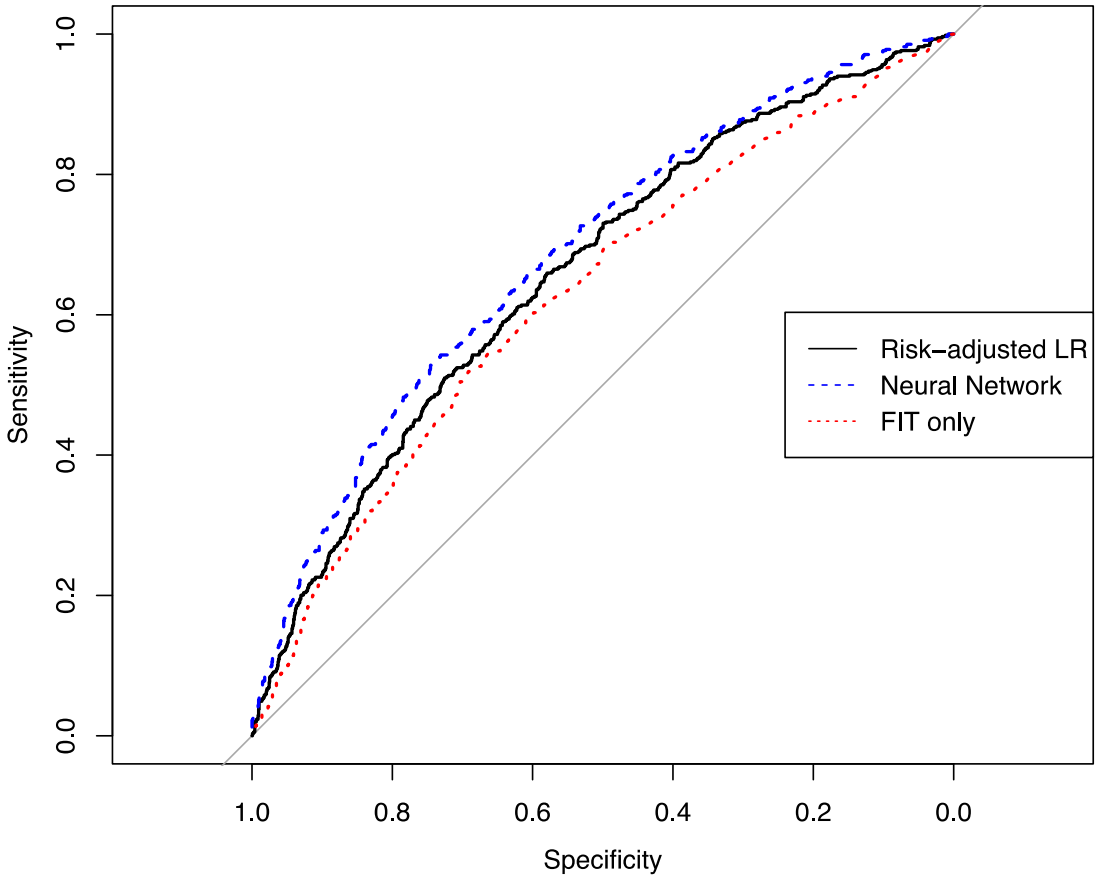


Figure 6: ROC curves for the final artificial neural network model compared to the risk-adjusted logistic regression model and FIT only. AUC (95% CI) for the Neural Network Model: 0.686 (0.659 - 0.712); AUC (95% CI) for the Risk-adjusted Logistic Regression Model: 0.659 (0.632 - 0.686); AUC (95% CI) for the FIT only: 0.628 (0.600 - 0.656).

Calibration

Calibration for the ANN assessed using the Hosmer-Lemeshow statistic gives a similar result (0.892) to the risk adjusted logistic regression model (0.898) indicating good model fit (See **Figure 7** for calibration plots).

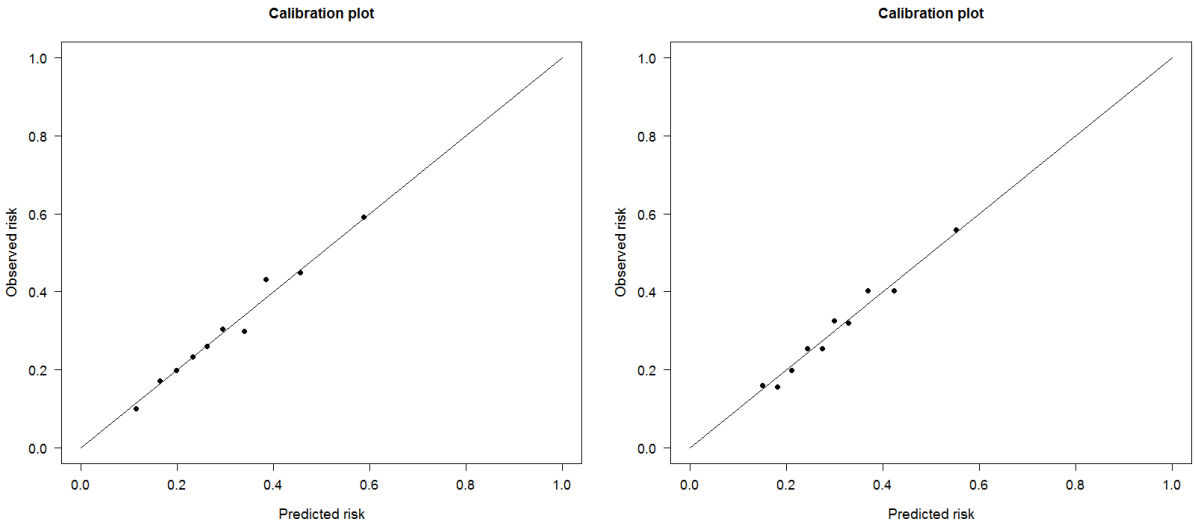


Figure 7: Calibration plots for the refined neural network y = 1.0033x (left) and the logistic regression model reported in the previous chapter.

As the number of group divisions used for the Hosmer-Lemeshow test effects the corresponding p-value, different group splits were investigated (5 to 15). On average the calibration was higher with the ANN across different risk group divisions compared with the risk-adjusted logistic regression model (See **Appendix 3** for results).

3.4 Test Accuracy

A two by two table is presented below for the ANN, risk-adjusted logistic regression model and FIT only, using a threshold of 160 µg/g and equivalent risk thresholds (**Table 5)**. Two by two tables were produced for thresholds between 30 to 180 µg/g (**Appendix 4**). At all thresholds, the sensitivity and specificity were greater for the ANN when compared with the logistic regression model (see **Table 6**). Focusing on a threshold of 160 µg/g which is the anticipated NHS BCSP cutpoint; the ANN has a sensitivity of 35.15% and a specificity of 85.57% compared to a sensitivity of 33.15% and specificity of 84.69% for the equivalent logistic regression model.

Applying the neural network at a threshold of 160 µg/g led to 24 more advanced adenomas being detected and the same number of cancers (30 more high risk adenomas and 6 less intermediate adenomas) compared with FIT only. Compared with the risk-adjusted logistic

regression model, 11 additional advanced adenomas were detected (13 more high risk adenomas, 2 less intermediate adenomas). The neural network therefore improved the diagnostic yield of high-risk adenomas as seen with the logistic regression model.

2 by 2 table for the neural network, the risk-adjusted logistic regression model and FIT only							
160 µg Hb/g faeces Threshold	Diagnostic Positive			Diagnostic Negative			Total
	FIT	Risk-adjusted	Neural Network	FIT	Risk-adjusted	Neural Network	
FIT/Risk Positive	**169** *37 - Cancer* *66 - High Risk Adenoma* *66 - Intermediate Risk Adenoma*	**182** *37 - Cancer* *83- High Risk Adenoma* *62 - Intermediate Risk Adenoma*	**193** *37 - Cancer* *96 - High Risk Adenoma* *60 - Intermediate Risk Adenoma*	**206** *70 - Abnormal* *92 - Low Risk Adenoma* *44 - Normal (No Abnormalities Found)*	**193** *69 - Abnormal* *81 - Low Risk Adenoma* *43 - Normal (No Abnormalities Found)*	**182** *62 - Abnormal* *79 - Low Risk Adenoma* *41 - Normal (No Abnormalities Found)*	**375**
FIT/Risk Negative	**380** *36 - Cancer* *148 - High Risk Adenoma* *196 - Intermediate Risk Adenoma*	**367** *36 - Cancer* *131 - High Risk Adenoma* *200 - Intermediate Risk Adenoma*	**356** *36 - Cancer* *118 - High Risk Adenoma* *202 - Intermediate Risk Adenoma*	**1055** *396 - Abnormal* *439 - Low Risk Adenoma* *220 - Normal (No Abnormalities Found)*	**1068** *397 - Abnormal* *450 - Low Risk Adenoma* *221 - Normal (No Abnormalities Found)*	**1079** *404 - Abnormal* *452 - Low Risk Adenoma* *223 - Normal (No Abnormalities Found)*	**1435**
Total	**549**			**1261**			**1810**
FIT only: Sensitivity 30.78%, Specificity 83.66%, PPV 45.07%, NPV 73.52%, FIT positivity 20.72%, Cancer Detection Rate 9.34% **Risk adjusted: Sensitivity 33.15%, Specificity 84.69%, PPV 48.53%, NPV 74.42%, FIT positivity 20.72%, Cancer Detection Rate 10.60%** **Neural Network: Sensitivity 35.15%, Specificity 85.57%, PPV 51.47%, NPV 75.19%, FIT positivity 20.72%, Cancer Detection Rate 10.66%**							

Table 5: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network. A threshold of 160 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.389 for the risk-adjusted model and 0.407 for the neural network. Profiles of outcome severity are also given. An 'Abnormal' result relates to other diagnoses such as haemorrhoids and inflammatory bowel diseases.

Model	FIT (µg Hb/g faeces)/ Risk Threshold (probability)	Sensitivity (%)	Specificity (%)
FIT only	30.00	88.34	22.20
Risk-adjusted LR	0.191	90.35	23.08
Neural Network	0.178	91.99	23.79
FIT only	40.00	76.68	38.94
Risk-adjusted LR	0.242	80.15	40.44
Neural Network	0.232	81.24	40.92
FIT only	50.00	69.03	50.04
Risk-adjusted LR	0.272	70.86	50.83
Neural Network	0.260	73.04	51.78
FIT only	60.00	60.66	59.24
Risk-adjusted LR	0.295	62.48	60.03
Neural Network	0.288	65.03	61.14
FIT only	70.00	55.19	64.63
Risk-adjusted LR	0.310	57.19	65.42
Neural Network	0.311	59.02	66.30
FIT only	80.00	51.18	69.31
Risk-adjusted LR	0.321	52.64	69.94
Neural Network	0.330	55.37	71.05
FIT only	90.00	45.72	72.56
Risk-adjusted LR	0.336	48.63	73.83
Neural Network	0.349	51.55	75.10
FIT only	100.00	42.44	75.26
Risk-adjusted LR	0.346	44.99	76.37
Neural Network	0.358	48.63	77.95
FIT only	110.00	40.07	77.08
Risk-adjusted LR	0.356	42.99	78.35
Neural Network	0.368	45.90	79.62
FIT only	120.00	38.07	78.59
Risk-adjusted LR	0.362	40.26	79.54
Neural Network	0.375	43.53	80.97
FIT only	130.00	34.79	80.33
Risk-adjusted LR	0.371	37.89	81.68
Neural Network	0.387	41.35	83.19
FIT only	140.00	33.70	81.60
Risk-adjusted LR	0.379	36.25	82.71
Neural Network	0.395	39.71	84.22
FIT only	150.00	32.42	82.39
Risk-adjusted LR	0.383	35.15	83.58
Neural Network	0.399	37.70	84.69
FIT only	160.00	30.78	83.66
Risk-adjusted LR	0.389	33.15	84.69
Neural Network	0.407	35.15	85.57
FIT only	170.00	29.87	84.30
Risk-adjusted LR	0.392	31.69	85.09
Neural Network	0.411	34.24	86.20
FIT only	180.00	28.60	85.57
Risk-adjusted LR	0.399	30.05	86.20
Neural Network	0.425	32.79	87.39

Table 6: Clinical sensitivity and specificity pairs for FIT thresholds between 30 and 180 µg Hb/g faeces and the corresponding risk thresholds.

3.5 Results Presented by Sex

The two by two tables are broken down further by sex in **Table 7** below for a threshold of 160 µg/g. Further thresholds are presented in **Appendix 4** and include results for; 30, 40, 50, 80, 150, 170 and 180 µg/g. In general, since males are at higher risk, the FIT detects more cancers/advanced adenomas in males compared to females. The risk-adjusted logistic regression model exacerbates this difference halving the number of females detected with high-risk adenomas. The neural network on the other hand levels out this difference in sex

by maintaining a similar number of advanced adenomas detected for females compared to using the FIT only.

The FIT result alone recalled 225 men (115 TP, 110 FP) of which 115 had cancer or advanced adenoma (51.11%), and 150 women (54 TP, 96 FP) of which 54 (36.00%) had cancer or advanced adenoma. The ANN recalled 279 men (146 TP, 133 FP) of which 146 (52.33%) had cancer or advanced adenoma, and 96 women (47 TP, 49 FP) of which 47 (48.96%) had cancer or advanced adenoma. The logistic regression model recalled 314 men (156 TP, 158 FP) of which 156 (49.68%) had cancer or advanced adenoma, and 61 women (26 TP, 35 FP) of which 26 (42.62%) had cancer or advanced adenoma. The ANN therefore when compared to the FIT result alone and logistic regression model improves the percentage of cancers/advanced adenomas detected in those recalled for further diagnostic tests (PPV). In addition, the difference between males and females in terms of cancers/advanced adenomas detected in those referred is reduced.

Both models decrease the number of false negatives for males compared to FIT only, with a greater reduction seen with the logistic regression model. Although both models increase the number of false negative results for women, this increase is greater with the logistic regression model compared to the ANN. For the false positive results, an increase is seen with both the logistic regression model and ANN compared with FIT only but a greater increase in this number is seen with the logistic regression model. For females on the other hand, the number of false positives is approximately halved for both models with a greater reduction seen with the logistic regression model.

For both models, there is an increase in the number of true positive results for males with a greater increase seen with the logistic regression model. There is also a decrease in the number of true positive results for women in both models but there are a greater number of TPs seen for the ANN, comparable to using FIT only.

The colorectal cancer/advanced adenoma detection rate for each model by screening history and sex subgroup is provided in **Appendix 5**. The ANN increases the cancer detection rate in female responders (4.85%) compared to the logistic regression model which decreases the cancer detection rate in this subgroup (1.88%) compared to the FIT alone (5.63%). When applying a model as a test in this way, it can lead to a change in the spectrum of diagnosed disease and in the subgroups where cancer is detected. This is a key element of Health Technology Assessment and can be considered relevant within this context.

2 by 2 table for FIT only, the risk-adjusted logistic regression model split by sex.														
160 µg Hb/g faeces Threshold		**Diagnostic Positive**							**Diagnostic Negative**						**Total**
		FIT only		**Risk-adjusted (LR)**		**Neural Network**			**FIT only**		**Risk-adjusted (LR)**		**Neural Network**		
		Male	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
FIT/Risk Positive	***Total***	**115**	**54**	**156**	**26**	**146**	**47**	***Total***	**110**	**96**	**158**	**35**	**133**	**49**	
	Cancer	27	10	29	8	27	10	***Low Risk Adenoma***	41	29	60	9	49	13	
	High risk Adenoma	45	21	72	11	71	25	***Abnormal***	51	41	66	15	59	20	**375**
	Intermediate risk Adenoma	43	23	55	7	48	12	***Normal (No Abnormalities Found)***	18	26	32	11	25	16	
FIT/Risk Negative		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
	Total	**243**	**137**	**202**	**165**	**212**	**144**	***Total***	**522**	**533**	**474**	**594**	**499**	**580**	
	Cancer	23	13	21	15	23	13	***Low Risk Adenoma***	222	174	203	194	214	190	
	High risk Adenoma	100	48	73	58	74	44	***Abnormal***	198	241	183	267	190	262	**1435**
	Intermediate risk Adenoma	120	76	108	92	115	87	***Normal (No Abnormalities Found)***	102	118	88	133	95	128	
Total		**549**							**1261**						**1810**

Table 7: 2 by 2 table for the neural network model, the risk adjusted logistic regression and FIT only split by sex. A threshold of 160 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.407 for the neural network and 0.389 for the risk-adjusted model. Profiles of outcome severity are also given.

3.6 Predictiveness Curve

The predictiveness curve is presented for FIT only, the logistic regression model and ANN in **Figure 8**. A risk threshold of 0.407 for the ANN and 0.389 for the logistic regression model is equivalent to a FIT cut-off of 160 µg Hb/g faeces. The FIT only model assigns greater risk to about 35% of participants (for those with predicted risks of around 0.2 and below) compared with the ANN and risk adjusted logistic regression model. Conversely the ANN assigns around 40% of individuals at higher probability (for those with predicted risks above 0.3) than both the LR and FIT only models. This greater difference seen with the ANN enhances the discrimination of the model to distinguish between those at higher risk and those at lower risk.

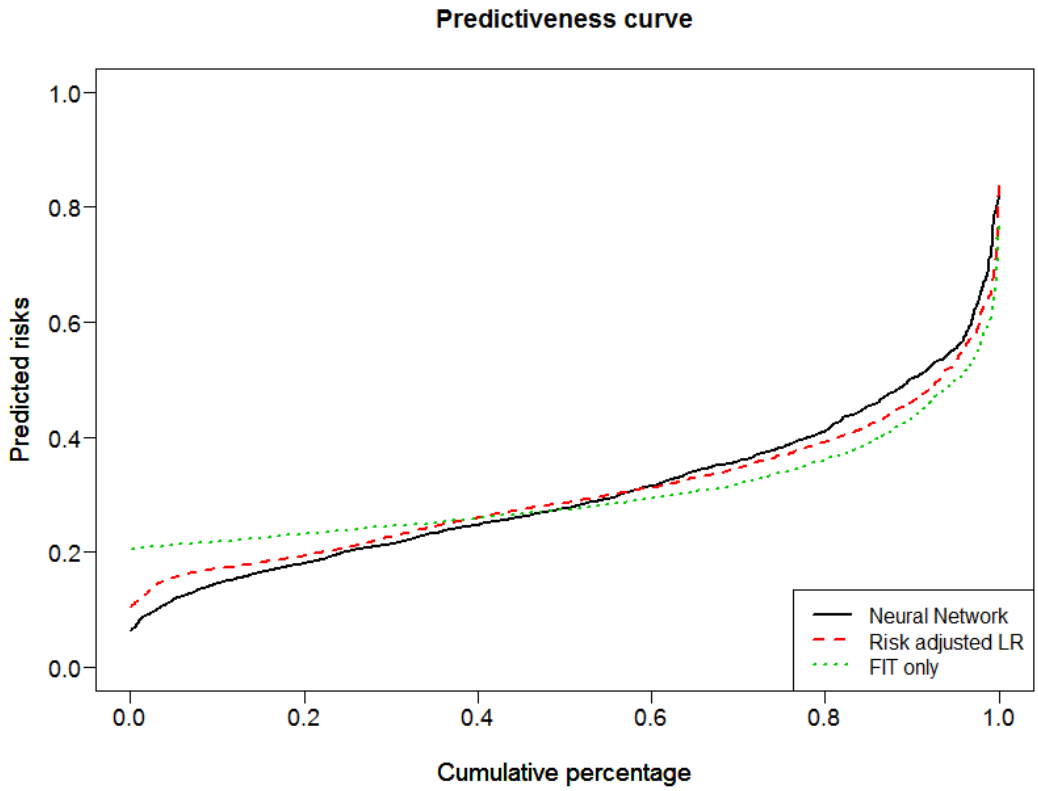


Figure 8: Predictiveness curve for the FIT only, logistic regression model (LR) and the artificial neural network (ANN). Predicted risk estimated from the different models versus the cumulative percentage of participants.

3.7 Patient Profiles for Each Model

Individual patient profiles with the corresponding risk probabilities for each model are presented in **Table 8** below. Referral decisions are based on a FIT threshold of 160 µg/g, which is equivalent to a risk threshold of 0.407 for the ANN and 0.389 for the LR model. For patient number 10 who is male, aged 62, a previous responder to FOBT screening, with a FIT result of 245.6 (well over the threshold to be referred using the FIT result alone) and an outcome at colonoscopy of 'Normal', the ANN assigns a probability of 0.389 which would mean that the individual would not be referred for additional diagnostic tests. Conversely, both the risk-adjusted logistic regression model and the FIT result alone would refer this individual for colonoscopy which carries its own risks. Looking at another example at the other end of the spectrum, patient number 9 who is male, aged 60, a previous non-responder to screening, with a FIT result of 33.4 (well under the threshold to be referred using the FIT result alone) and an outcome of 'high-risk adenoma', the ANN assigns a probability of 0.673 which would mean that this individual would be referred for additional tests. On the other hand, the risk-adjusted logistic regression model and the FIT result alone would assign this as a negative test result (false negative) which would miss this high risk diagnosis. Further examples are presented in the table along with the IMD score for reference, although this was not found to be significant in the logistic regression model and did not improve cross-validated deviance for the ANN.

Patient Profile	Age	Sex	FIT Result	Screening History	IMD Score	Outcome	ANN probability	LR probability	FIT only probability
1	66	Male	169	Previous responder	48.18	Low-risk Adenoma	0.385	0.401*	0.363*
2	60	Female	167	First Time Screenee	5.04	Low-risk Adenoma	0.241	0.199	0.362*
3	62	Male	162.2	Previous responder	16.43	Abnormal	0.376	0.379	0.359*
4	68	Female	161.4	Previous responder	40.95	Low-risk Adenoma	0.371	0.282	0.359*
5	71	Female	157.8	Previous Responder	19.35	Cancer	0.453*	0.292	0.357
6	69	Male	157.4	Previous Responder	9.06	Cancer	0.518*	0.410*	0.356
7	68	Male	155.6	Previous Responder	51.43	High-risk Adenoma	0.486*	0.404*	0.355
8	72	Female	154.4	Previous Responder	2.53	Intermediate-risk Adenoma	0.449*	0.295	0.355
9	60	Male	33.4	Previous Non responder	7.55	High Risk Adenoma	0.673*	0.358	0.238
10	62	Male	245.6	Previous Responder	41.72	Normal (No abnormalities found)	0.389	0.414*	0.395*

Table 8: Patient Profiles for 10 individuals with the corresponding probabilities estimated from the artificial neural network (ANN) and logistic regression models (LR) and for the FIT result only. A star '' next to the probability indicates that the individual would have been referred based on that model or FIT result using a FIT threshold of 160µg/g, which is equivalent to a risk threshold of 0.407 for the ANN and 0.389 for the LR model.*

The patient profiles give an idea about the relative importance of the variables in the ANN model which for the logistic regression model would be provided in the form of beta coefficients and corresponding odds ratios. The plot below uses Garson's algorithm to show the relative importance of the input variables for the ANN (**Figure 9**).[64] The relative magnitude of the screening history variables 'previous responder' and 'previous non-responder' compared to a first time invitee at baseline has the greatest effect on the model outcome, followed by age, FIT result and finally sex. The odds ratios in the previous chapter for the logistic regression model suggests that a 'previous non responder' has the greatest association with the outcome (OR= 2.271, CI: 1.422-3.667). The other variables are more difficult to interpret since a one-unit change in a continuous variable has an additive effect on the overall odds ratio.

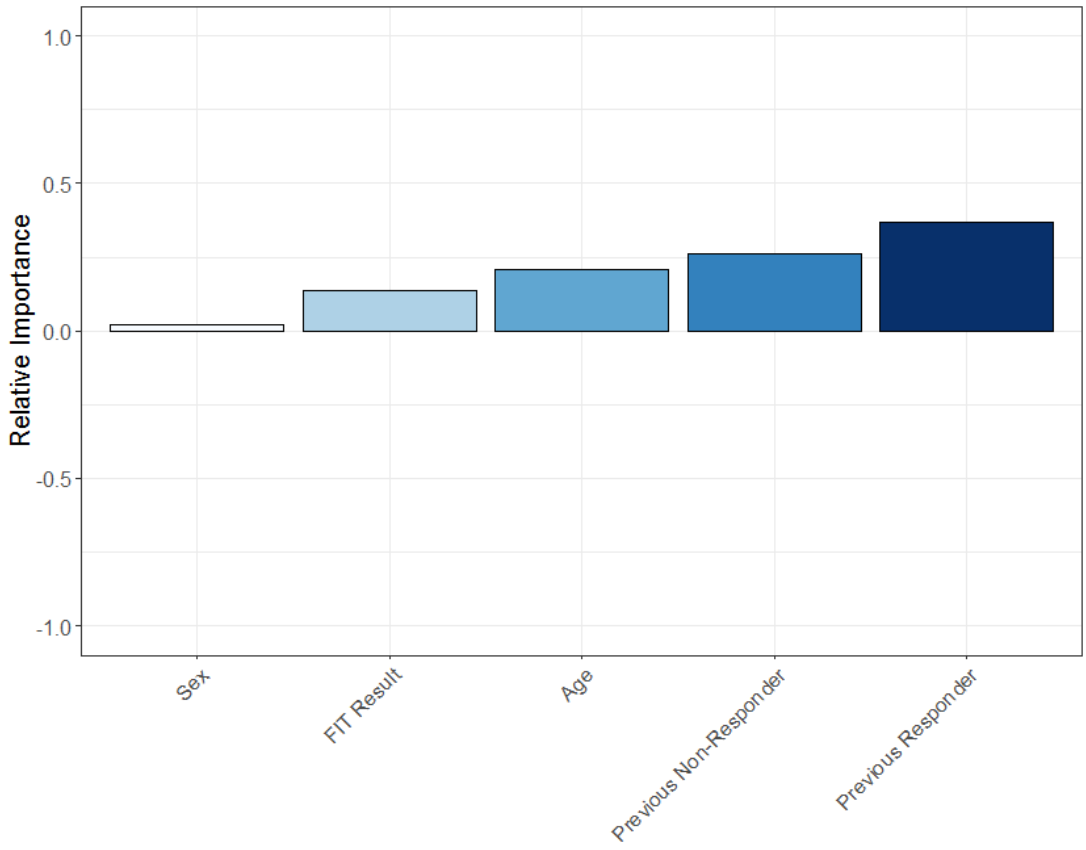


Figure 9: Plot using Garson's algorithm to show the relative importance of the input variables for the ANN.

4.0 DISCUSSION

4.1 Statement of Principal Findings

The optimal ANN used for comparison against the risk-adjusted logistic regression model consisted of 5 input nodes, 3 hidden layer nodes and 1 output node. Standardisation of continuous variables reduced the cross-validated deviance and allowed for greater generalizability. A weight decay of 0.01 was used for the model since this gave the lowest SSE (346.04). Removing 4 weight connections (8, 4, 16 and 17) improved the cross-validated deviance further (2077.694). This finalized model with 18 weights was used for comparison.

The discrimination of the ANN (0.69, CI: 0.66-0.71) was greater than the risk adjusted logistic regression model (0.66, CI: 0.63-0.69) and on average calibration was higher with the ANN across different group divisions. The sensitivity and specificity of the ANN was greater at all thresholds. At 160 µg/g, sensitivity was 35.15% compared to 33.15% for the logistic regression model and specificity was 85.57% compared to 84.69%.

At this threshold, the ANN detected 11 additional advanced adenomas (13 more high risk adenomas and 2 fewer intermediate adenomas) compared with the risk-adjusted logistic regression model and 24 more advanced adenomas (30 more high risk adenomas and 6 fewer intermediate adenomas) compared with the FIT alone. Based on the results from these data, for every 1,000,000 people invited to screening an estimate of an additional 586 cancers/advanced adenomas (4,715/1,000,000) would be detected compared to using the FIT alone (4,129/1,000,000).

When investigating the breakdown of referrals by sex in the 2 by 2 tables for a threshold of 160 µg Hb/g faeces, it was found that cancers and advanced adenomas were being detected more in referrals for men than women. The neural network equalized the difference in cancers/advanced adenomas detected between men and women seen with the logistic regression and FIT only models. At 160 µg Hb/g faeces, compared with FIT only the neural network increased the number of high-risk adenomas detected for women (cancers stayed the same) and halves the number of false positive results for women. Although the neural network recalls fewer women, the PPV is increased compared to the other models and is similar between the sexes (men – 52.33%; women – 48.96%).

At either end of the risk spectrum, the ANN helped to enhance the difference between those at low-risk of disease with those at higher risk, which allowed better discrimination

capabilities of the model/test. This difference was visually represented in the predictiveness curve.

Garson's algorithm was used to assess the relative importance of the input variables for the ANN. The relative importance was greatest for the screening history variables, 'previous responder' and 'previous non-responder', followed by age, FIT result and sex. The screening history variables also had the greatest odds ratio for the logistic regression model.

Examples of patient profiles were presented. One example showed that the ANN assigned a lower probability for an individual with a normal outcome at colonoscopy whereas the LR model and FIT only would have referred this individual for further unnecessary testing. The repercussions of increased false positive results are putting people through unnecessary diagnostic tests, which in the case of colonoscopy carries risk of bleeding and perforation of the bowel (which can lead to death). There is also a psychological effect for these individuals as they may think they have cancer or its precursor.

Another example of a patient profile is presented whereby they did not meet the referral threshold for both the LR and FIT only models but did so for the neural network when they had a 'high risk' diagnostic outcome. The number of false negative results should be limited as far as possible since this has the effect of missing a potential cancer/advanced adenoma and relates to the sensitivity of the test

4.2 Strengths and Weaknesses of the Study

This study used the same dataset as the previous chapter for developing an ANN risk prediction model. In addition, the same variables were used for the final model as the risk-adjusted logistic regression model allowing direct comparisons in terms of model performance and test accuracy. As described in the previous chapter, the pilot was implemented within a live screening programme providing good quality data for model development.

Different approaches were taken to improve the generalization of the model including investigating different values of the weight decay parameter which provided the lowest SSE, and pruning network connections. By reducing complexity in this way the generalization of the model is improved in external populations. Although the network was pruned, this was just based on absolute magnitude with the subsequent change in cross-

validated deviance monitored as each weight was dropped. Conversely, an optimization algorithm which tried every single iteration of weight connection removal could be used. For example, more formal methods of pruning weights have been formalised in the literature; optimal brain damage, optimal brain surgery and genetic algorithms have been applied to subset selection.[15 92-95] The optimal brain damage and optimal brain surgery methods are variants of the Wald (or likelihood ratio) test. The method used for this study however dropped all the small magnitude weights to achieve a model with a significantly lower cross-validated deviance compared to the logistic regression model. Furthermore, for ANN model development, there are empirical and methodological issues which remain to be resolved,[1] there are fewer guidelines to be followed within the literature.

Finally, the same limitations to the dataset as described in the previous chapter in relation to follow up data and interval cancers not being recorded on BCSS and the participant flow have been inherited in this study.

4.3 Strengths and weaknesses in relation to other studies

As far as can be identified, neural networks have not been developed for use in an average risk screening population for screening referral decisions. ANNs have however been developed for diagnosis, survival prediction, cancer relapse and distant metastases for colorectal cancer.[37 96]

As identified in the methodology review by Dreiseitl and Ohno-Machado,[9] the model building procedure was reported more often in logistic regression based models compared with ANNs based on variable selection, parameter selection and over-fitting avoidance. This research fully reports the construction of the network architecture and the selection of the weight decay parameter value as well as the step by step removal of weights during the network pruning to allow the model to be reproduced. The importance of reporting the model-building procedure is reflected in the TRIPOD guidelines under item 10b; 'Specify type of model, all model-building procedures (including any predictor selection), and methods for internal validation'.[97] Many studies only report the model with best discrimination which may reflect a model which is over-fitted with over-optimistic model performance and therefore has worse performance when applied in external populations. Model selection is driven by both the data and clinical context and this needs to be fully reported and justified within the study.

4.4 Practical Implications

Although the performance of the neural network was better than the logistic regression model, as discussed in the introduction there are several barriers which have hindered the widespread practical application of machine learning approaches.[32 50-53] This study focused on providing a reportable equation which would allow further researchers to assess and validate the model and to enhance computational transportability since it could be applied in practice with minimal software requirements. To try and open the 'black box', this study focused on using additional methods to present the relative importance of variables in the form of Garson's algorithm, patient profiles and predictiveness curves to enhance model interpretability. External validation and impact studies would however be required to assess the effect on patient outcomes and to determine acceptability to patients and clinicians. Assessing outcomes such as these will improve clinician 'trust' in the use of machine learning algorithms.

Both models give the absolute risk prediction for each individual and this can be used to make clinical decisions regarding screening referral by setting an appropriate 'risk threshold'. The BCSS has capacity within the database to use a risk algorithm for screening referral. The model equation is fully reported in this study and could be used for subsequent external validation and future impact studies. The methods for assessing external validation of neural networks would however need to be further explored but can be based on the methods used for a logistic regression validation if using a logistic activation and output function. Furthermore, if additional predictors are included in the model in the future, non-linear predictors and model interactions may be better captured with a neural network or other machine-learning algorithm.

The FIT detected more cancers/advanced adenomas in males compared to females and this difference was exacerbated when applying the risk-adjusted logistic regression model. The neural network on the other hand levels out this difference in sex by increasing the number of high-risk adenomas detected for women (cancers stay the same) and halving the number of false positive results for women (threshold 160 µg Hb/g faeces). Depending on the screening programme aims, by referring a greater number of males this may cause potential issues in the screening community. A further method to combat this difference in the sexes is to consider separate algorithms for males and females.

When applying a model as a test in this way, it can lead to a change in the spectrum of diagnosed disease and in the subgroups where cancer is detected. This is a key element of Health Technology Assessment and requires further investigation in this area.

4.5 Future Research

The risk based models developed in the current and previous chapters could be refined further by utilising additional risk predictors available from the BCSS as described in the previous chapter as well as using follow up information once the FIT is rolled out in 2018. This refined model could then be assessed in a further dataset using FIT for external validation (for instance, Scotland have implemented the FIT, along with the Isle of Man).

The impact of the risk prediction model could be investigated by recalling an individual if either the FIT result alone or the risk-based model suggests cancer could be detected at colonoscopy and assessing the corresponding diagnostic accuracy and patient outcomes.

The BCSS has an inbuilt function of using 1/n data for screening participants. For the pilot, 1 out of every 28 invitations was assigned a FIT, a similar approach here could be used to assess risk adjusted screening or to assign a range of thresholds so data can be retained for future analysis.

Although the performance of the neural network is significantly better than using a logistic regression model, another approach to improving discrimination power is to include a richer set of predictors. As identified in the previous chapter, further predictors could be obtained from the BCSS such as previous FIT results once this screening test is implemented in England, as well as colonoscopy and flexible sigmoidoscopy results. Alternatively, studies identified in the systematic review suggest that discrimination can be improved by incorporating the results of further lab tests,[98-100] or lifestyle parameters.[101 102] A study identified from the systematic review which combines both lifestyle information with lab based parameters in an accelerated failure time model shows enhanced discrimination (ROC AUC: 0.86, 95% CI – 0.85, 0.87).[101] Another potential data source for richer predictors such as these which may improve discriminatory power includes electronic GP records which contain a wide array of information including; symptoms, prescriptions, lab test results and lifestyle parameters. The BCSS receives data for its screening participants from the NHS Spine[103] which houses demographic information (name, address, postcode, NHS number, date of birth for those aged 60-74) drawn from GP records. There is capacity therefore to draw further information from the spine or from GP

records to improve screening referral decisions. Risk prediction studies for colorectal cancer screening have suggested the possible use of electronic GP records to provide this extra risk information.[102 104]

Machine learning approaches have less guidance over more conventional modelling techniques such as logistic regression. The TRIPOD guidelines focus mainly on regression models but some of the principles are equally valid to machine learning methods. ANNs are less well reported compared to logistic regression as found in systematic reviews.[17 37] There are also empirical and methodological issues which remain to be resolved,[1] and fewer guidelines in the literature. Therefore future research in prediction modelling should focus on guidelines for such models to improve reporting and transparency of results. This may also have a corresponding effect on their clinical use in practice.

5.0 CONCLUSIONS

Although it is often argued that neural networks are more difficult to interpret and are often likened to a 'black box' this result shows the promise of machine learning algorithms for use in screening decisions and clinical practice. Both the logistic regression and neural network can give the absolute risk for each individual and this can be used for screening referral decisions by setting an appropriate 'risk threshold'. With the shift to larger and more complex electronic health data, machine-learning algorithms may be better placed to deal with larger amounts of data and non-linear associations when compared with conventional models such as logistic regression.

Another approach to improving model performance and discrimination power is to consider a richer set of predictors. The BCSS receives data for its participants from the NHS Spine which houses demographic information (name, address, postcode, NHS number, date of birth for those aged 60-74) drawn from GP records. There is capacity to draw further information from the Spine or from GP records to improve screening referral decisions. For instance, lab based parameters and lifestyle parameters were shown to improve discrimination when combined with FIT in risk prediction models in Chapter 2. A study identified from the review combined both lifestyle information and lab results with the FIT in an accelerated failure time model showing enhanced discrimination.[101] Further to this, several studies based on primary care data have investigated predictors which could be used in a prediction model for primary care referral.[5 105 106] The next chapter will investigate the use of an anonymised GP record database (THIN) to define a screening

population, determine how complete potential predictors are for this population and to develop risk prediction models (using survival analysis methods) for use in screening referral decisions by combining richer predictor information. This will help to identify potential further predictors which may enhance a risk based screening model.

6.0 REFERENCES

1. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology. 1996;49(11):1225-31.
2. Adams ST, Leveson SH. Clinical prediction rules. BMJ (Clinical research ed). 2012;344:d8312.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.
4. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989;81(24):1879-86.
5. Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut. 2011;60(9):1242-8.
6. Kidney E, Berkman L, Macherianakis A, Morton D, Dowswell G, Hamilton W, et al. Preliminary results of a feasibility study of the use of information technology for identification of suspected colorectal cancer in primary care: the CREDIBLE study. Br J Cancer. 2015;112(s1):S70-S6.
7. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. The British journal of general practice : the journal of the Royal College of General Practitioners. 2012;62(594):e29-37.
8. Ayer T, Chhatwal J, Alagoz O, Kahn CE, Jr., Woods RW, Burnside ES. Informatics in radiology: comparison of logistic regression and artificial neural network models in breast cancer risk estimation. Radiographics : a review publication of the Radiological Society of North America, Inc. 2010;30(1):13-22.
9. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.
10. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015;162(1):W1-73.
11. Chen S-T, Hsiao Y-H, Huang Y-L, Kuo S-J, Tseng H-S, Wu H-K, et al. Comparative Analysis of Logistic Regression, Support Vector Machine and Artificial Neural Network for the Differential Diagnosis of Benign and Malignant Solid Breast Tumors by the Use of Three-Dimensional Power Doppler Imaging. Korean Journal of Radiology. 2009;10(5):464-71.
12. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics. 2001;34(1):28-36.
13. Gurney K. An Introduction to Neural Networks. London: UCL Press; 1997.
14. Ripley BD. Statistical Data Mining 2002 [cited 2016 17th November]. Available from: https://www.stats.ox.ac.uk/pub/bdr/SDM2002/DM2002.pdf.
15. Ripley BD. Pattern Recognition and Neural Networks. Cambridge: Cambridge University Press; 2007.
16. Spackman KA. Maximum likelihood training of connectionist models: comparison with least squares back-propagation and logistic regression. Proceedings Symposium on Computer Applications in Medical Care. 1991:285-9.

17. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8 Suppl):1636-42.
18. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. The Lancet. 1995;346(8982):1075-9.
19. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. Cancer. 2001;91(8 Suppl):1615-35.
20. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European heart journal. 2017;38(23):1805-14.
21. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.
22. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. Asian Pacific journal of cancer prevention : APJCP. 2018;19(7):1747-52.
23. Yassin NIR, Omran S, El Houby EMF, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Computer methods and programs in biomedicine. 2018;156:25-45.
24. Bertolaccini L, Solli P, Pardolesi A, Pasini A. An overview of the use of artificial neural networks in lung cancer research. Journal of thoracic disease. 2017;9(4):924-31.
25. Nishio M, Sugiyama O, Yakami M, Ueno S, Kubo T, Kuroda T, et al. Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. PloS one. 2018;13(7):e0200721-e.
26. Hu X, Cammann H, Meyer HA, Miller K, Jung K, Stephan C. Artificial neural networks and prostate cancer--tools for diagnosis and management. Nature reviews Urology. 2013;10(3):174-82.
27. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology. 2017;2(4):230-43.
28. Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. The Lancet Psychiatry. 2016;3(1):13-5.
29. Wong T, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA. 2016;316(22):2366-7.
30. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama. 2016;316(22):2402-10.
31. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015;13:8-17.
32. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clinical chemistry and laboratory medicine. 2018;56(4):516-24.
33. Yang HJ, Cho CW, Kim SS, Ahn KS, Park SK, Park DI. Application of deep learning to predict advanced neoplasia using big clinical data in colorectal cancer screening of asymptomatic adults. Gastrointestinal endoscopy. 2018.
34. Carin L, Pencina MJ. On deep learning for medical image analysis. JAMA. 2018;320(11):1192-3.
35. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016;375(13):1216-9.
36. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. Neural networks : the official journal of the International Neural Network Society. 2006;19(4):408-15.

37. Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular Cancer. 2005;4:29-.
38. Biglarian A, Bakhshi E, Gohari MR, Khodabakhshi R. Artificial neural network for prediction of distant metastasis in colorectal cancer. Asian Pacific journal of cancer prevention : APJCP. 2012;13(3):927-30.
39. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015;13:1.
40. Betes M, Munoz-Navas MA, Duque JM, Angos R, Macias E, Subtil JC, et al. Use of colonoscopy as a primary screening test for colorectal cancer in average risk people. Am J Gastroenterol. 2003;98:2648-54.
41. Williams TG, Cubiella J, Griffin SJ, Walter FM, Usher-Smith JA. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC gastroenterology. 2016;16(1):63.
42. Hamilton W, Round A, Sharp D, Peters TJ. Clinical features of colorectal cancer before diagnosis: a population-based case-control study. Br J Cancer. 2005;93(4):399-405.
43. Mahadavan L, Loktionov A, Daniels IR, Shore A, Cotter D, Llewelyn AH, et al. Exfoliated colonocyte DNA levels and clinical features in the diagnosis of colorectal cancer: a cohort study in patients referred for investigation. Colorectal Disease: The Official Journal Of The Association Of Coloproctology Of Great Britain And Ireland. 2012;14(3):306-13.
44. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. British Journal of Cancer. 2009;101(Suppl 2):S80-S6.
45. Hamilton W, Green T, Martins T, Elliott K, Rubin G, Macleod U. Evaluation of risk assessment tools for suspected cancer in general practice: a cohort study. The British journal of general practice : the journal of the Royal College of General Practitioners. 2013;63(606):e30-6.
46. Usher-Smith JA, Walter FM, Emery JD, Win AK, Griffin SJ. Risk Prediction Models for Colorectal Cancer: A Systematic Review. Cancer Prev Res (Phila). 2016;9(1):13-26.
47. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biometrical journal Biometrische Zeitschrift. 2014;56(4):588-93.
48. Obermeyer Z, Lee TH. Lost in Thought — The Limits of the Human Mind and the Future of Medicine. New England Journal of Medicine. 2017;377(13):1209-11.
49. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.
50. Berner ES, Ozaydin B. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2353-4.
51. Cabitza F, Rasoini R, Gensini G. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8.
52. Huesch MD. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2355-6.
53. McDonald L, Ramagopalan SV, Cox AP, Oguz M. Unintended consequences of machine learning in medicine? F1000Research. 2017;6:1707-.
54. Povyakalo AA, Alberdi E, Strigini L, Ayton P. How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. Medical decision making : an international journal of the Society for Medical Decision Making. 2013;33(1):98-107.
55. Licitra L, Trama A, Hosni H. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2354-.
56. Lasko TA, Walsh CG, Malin B. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2355-.

57. Simon R. Class probability estimation for medical studies. Biometrical journal Biometrische Zeitschrift. 2014;56(4):597-600.
58. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biometrical journal Biometrische Zeitschrift. 2014;56(4):601-6.
59. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014;14(1):137.
60. Harrell Jr FE. Is Medicine Mesmerized by Machine Learning? 2018 [Available from: http://www.fharrell.com/post/medml/.
61. Binder H. What subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction? Biometrical journal Biometrische Zeitschrift. 2014;56(4):584-7.
62. Kruppa J, Liu Y, Biau G, Kohler M, Konig IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. Biometrical journal Biometrische Zeitschrift. 2014;56(4):534-63.
63. Kruppa J, Liu Y, Diener HC, Holste T, Weimar C, Konig IR, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. Biometrical journal Biometrische Zeitschrift. 2014;56(4):564-83.
64. Garson GD. Interpreting neural-network connection weights. AI Expert. 1991;6(4):46-51.
65. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ (Clinical research ed). 2013;346:e5595.
66. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ (Clinical research ed). 2013;346:e5793.
67. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS medicine. 2013;10(2):e1001380.
68. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2):e1001381.
69. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11).
70. Moss S, Mathews C, Day TJ, Smith S, Seaman HE, Snowball J, et al. Increased uptake and improved outcomes of bowel cancer screening with a faecal immunochemical test: results from a pilot study within the national screening programme in England. Gut. 2016.
71. Fraser CG, Allison JE, Halloran SP, Young GP. A proposal to standardize reporting units for fecal immunochemical tests for hemoglobin. Journal of the National Cancer Institute. 2012;104(11):810-4.
72. Department of Health. NHS public health functions agreement 2015-16. Service specification no.26 Bowel Cancer Screening Programmme 2014 [cited 2016 17th November]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/383200/1516_No26_NHS_Bowel_Cancer_Screening_Programme_Final.pdf.
73. Winawer SJ, Zauber AG. The advanced adenoma as the primary target of screening. Gastrointestinal endoscopy clinics of North America. 2002;12(1):1-9, v.

74. Brenner H, Hoffmeister M, Stegmaier C, Brenner G, Altenhofen L, Haug U. Risk of progression of advanced adenomas to colorectal cancer by age and sex: estimates based on 840 149 screening colonoscopies. Gut. 2007;56(11):1585-9.
75. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
76. Hadley Wickham and Romain Francois. dplyr: A Grammar of Data Manipulation. R package version 0.4.1 ed2015.
77. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. European journal of epidemiology. 2011;26(4):261-4.
78. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12:77.
79. Venables WNR, B. D. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.
80. Beck M. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. R package version 1.3.1 ed2015.
81. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009.
82. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-13 ed2014.
83. Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods. 2000;43(1):3-31.
84. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer; 2009.
85. Reed R. Pruning algorithms-a survey. IEEE transactions on neural networks. 1993;4(5):740-7.
86. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R New York: Springer; 2013.
87. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. . Communications in Statistics. 1980;A10:1043-69.
88. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. Statistics in Medicine. 2013;32(1):67-80.
89. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
90. Schreuders EH, Ruco A, Rabeneck L, Schoen RE, Sung JJY, Young GP, et al. Colorectal cancer screening: a global overview of existing programmes. Gut. 2015.
91. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the Predictiveness of a Marker with Its Performance as a Classifier. American Journal of Epidemiology. 2008;167(3):362-8.
92. Tolstrup N. Pruning of a large network by optimal brain damage and surgeon: an example from biological sequence analysis. International journal of neural systems. 1995;6(1):31-42.
93. Yang J, Honavar V. Feature Subset Selection Using a Genetic Algorithm. In: Liu H, Motoda H, editors. Feature Extraction, Construction and Selection: A Data Mining Perspective. Boston, MA: Springer US; 1998. p. 117-36.
94. Montana DJ, Davis L, editors. Training Feedforward Neural Networks Using Genetic Algorithms. IJCAI; 1989.
95. LeCun Y, Denker JS, Solla SA, Howard RE, Jackel LD, editors. Optimal brain damage. NIPs; 1989.

96. Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PW, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. Lancet (London, England). 1997;350(9076):469-72.
97. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Annals of internal medicine. 2003;138(1):W1-12.
98. Karl J, Wild N, Tacke M, Andres H, Garczarek U, Rollinger W, et al. Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. Clinical Gastroenterology & Hepatology. 2008;6(10):1122-8.
99. Kim BC, Joo J, Chang HJ, Yeo HY, Yoo BC, Park B, et al. A predictive model combining fecal Calgranulin B and fecal occult blood tests can improve the diagnosis of colorectal cancer. PloS one [Internet]. 2014; 9(9 // () *National Cancer Center* // () *National Cancer Center*). Available from: http://onlinelibrary.wiley.com/o/cochrane/clcentral/articles/537/CN-01014537/frame.html http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0106182&type=printable.
100. Tao S, Haug U, Kuhn K, Brenner H. Comparison and combination of blood-based inflammatory markers with faecal occult blood tests for non-invasive colorectal cancer screening. British Journal of Cancer. 2012;106(8):1424-30.
101. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. International journal of cancer Journal international du cancer. 2014;135(5):1203-12.
102. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
103. NHS Digital. Spine 2017 [Available from: https://digital.nhs.uk/spine.
104. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014.
105. NICE. Referral guidelines for suspected cancer 2005 [Available from: http://www.nice.org.uk/nicemedia/live/10968/29814/29814.pdf.
106. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer. 2009;101 Suppl 2:S80-6.

7.0 APPENDICES

Appendix 1: R scripts used for model development and to assess performance

```
###Artificial Neural Network###
# ------------------------------------------------------------------------------------------------------------
# Read Data in for modelling
ccFIT20 <- read.csv("FITroutine.csv")
# ------------------------------------------------------------------------------------------------------------
#Load package and library for nnet (in base R) and further packages
install.packages("nnet")
library(nnet)
install.packages("NeuralNetTools")
library("NeuralNetTools")
install.packages("ggplot2")
library("ggplot2")
install.packages("pROC")
library("pROC")
# ------------------------------------------------------------------------------------------------------------
# Standardisation of continuous variables

ccFIT20$stan.FIT <- (log(ccFIT20$TRANS_FIT_KIT_RESULT +1) - mean(log(ccFIT20$TRANS_FIT_KIT_RESULT
+1)))/sd(log(ccFIT20$TRANS_FIT_KIT_RESULT +1))
mean(ccFIT20$stan.FIT)

ccFIT20$stan.FIT2 <- ccFIT20$TRANS_FIT_KIT_RESULT -
(mean(ccFIT20$TRANS_FIT_KIT_RESULT)/sd(ccFIT20$TRANS_FIT_KIT_RESULT))
mean(ccFIT20$stan.FIT2)


ccFIT20$stan.age <- (ccFIT20$AGE_AT_EPISODE_START -
mean(ccFIT20$AGE_AT_EPISODE_START))/sd(ccFIT20$AGE_AT_EPISODE_START)
mean(ccFIT20$stan.age)

ccFIT20$stan.IMD <- (ccFIT20$IMD_SCORE - mean(ccFIT20$IMD_SCORE))/sd(ccFIT20$IMD_SCORE)
mean(ccFIT20$stan.IMD)
# ------------------------------------------------------------------------------------------------------------
#Set seed for replication of results

set.seed(23456)

neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
linout = FALSE, decay = 0.0, entropy = TRUE, maxit = 500)

#Print structure of network
summary(neuralnetmod)

# ------------------------------------------------------------------------------------------------------------
#Repeat the above for different weight decay values, hidden layer nodes to determine cross validated deviance and therefore
network architecture of best fitting neural network model
# ------------------------------------------------------------------------------------------------------------

# deviance and weight decay
nnet.deviance <- function(nnet.fitted){
  nnet.obs <- as.integer(nnet.fitted$residuals + nnet.fitted$fitted.values)
  e.pos <- nnet.obs * log(nnet.obs / nnet.fitted$fitted.values)
  e.neg <- (1 - nnet.obs) * log((1 - nnet.obs)/(1 - nnet.fitted$fitted.values))
  e.all <- rowSums(data.frame(e.pos, e.neg), na.rm = TRUE)
  E.mod <- sum(e.all)
  penalty <- sum(nnet.fitted$wts ^ 2) * nnet.fitted$decay
  list(deviance = 2* E.mod, E = E.mod, penalty = penalty, E.crit = E.mod + penalty)
}
nnet.deviance(nnet.fitted = neuralnetmod)

# ------------------------------------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------------------------------------
#To determine cross validated deviance using 10 fold cross validation

CVnn <- function(mod, m.dat, nreps=10, verbose = TRUE, seedno = 1234, ...)
{
 pred.cv <- vector(length = dim(m.dat)[1])
 set.seed(seedno)
 for(rep in 1:nreps) {
  rand <- sample(10, dim(m.dat)[1], replace = TRUE)
  if (verbose == TRUE){
   cat("Replication =",rep,"\n")
   cat("Randomisation ",rand[1:10],". . .","\n")
   cat("Size",paste(mod$n, collapse = "-"),"and decay ",mod$decay,"\n")
  }
  for (i in sort(unique(rand))) {
   if(verbose == TRUE){cat("fold ", i, "\n", sep="")}
   cvmod.nn <- update(mod, data = m.dat[rand != i, ], trace = FALSE)
   pred.cv[rand == i] <- pred.cv[rand == i] +
    predict(cvmod.nn, m.dat[rand == i,])
  }
 }
 nnet.obs <- as.integer(mod$residuals + mod$fitted.values)
 e.pos <- nnet.obs * log(nnet.obs / (pred.cv / nreps))
 e.neg <- (1 - nnet.obs) * log((1 - nnet.obs)/(1 - (pred.cv / nreps)))
 e.all <- rowSums(data.frame(e.pos, e.neg), na.rm = TRUE)
 E.mod <- sum(e.all)
 cv.deviance <- 2 * E.mod
 return(cv.deviance)
}
#Insert neural network model name
CVnn(mod = neuralnetmod, m.dat = ccFIT20, nreps = 10, seedno = 51234, verbose = FALSE)

# ------------------------------------------------------------------------------------------------------------
#5-3-1 gives lowest cross validated deviance at a decay of 0 and 3 hidden layer nodes
# ------------------------------------------------------------------------------------------------------------
#To visualise the ANN

set.seed(23456)
neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
linout = FALSE, decay = 0.0, entropy = TRUE, maxit = 500)
summary(neuralnetmod)
plotnet(neuralnetmod, cex_val =0.8)

#For numerical input
wts <- neuralweights(neuralnetmod)
struct <- wts$struct
wts <- unlist(wts$wts)
# plot
plotnet(wts, struct = struct)
# ------------------------------------------------------------------------------
# To obtain predicted probabilties from the ANN
neuralnetmod$fitted.values
# ------------------------------------------------------------------------------------------------------------
####Optimising the selected model###
# ------------------------------------------------------------------------------------------------------------
#Determining weight decay

#Investigating different weight decay values and their effect on SSE from 0.0 to 1.0

error = vector("numeric", 100)
wdecay = seq(0.0, 1.0, length.out=100)

for(i in 1:100) {
 set.seed(23456)
 fit <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
         linout = FALSE, decay = wdecay[i], entropy = TRUE, maxit = 500)
 error[i] <- sum(fit$residuals^2)
}
plot(wdecay, error)

errordecay <- data.frame(wdecay, error)
```

```
#Find the minimum SSE
min(errordecay$error)


#Investigate different weight decay values and their effect on SSE from 0.0001 to 0.1
#SSE1 = sum(fitnn1$residuals^2)

error = vector("numeric", 100)
wdecay = seq(0.0001, 0.1, length.out=100)


for(i in 1:100) {
  set.seed(23456)
  fit <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
              linout = FALSE, decay = wdecay[i], entropy = TRUE, maxit = 500)
  error[i] <- sum(fit$residuals^2)
}
plot(wdecay, error)

errordecay <- data.frame(wdecay, error)

#Find the minimum SSE
min(errordecay$error)

# ------------------------------------------------------------------------------------------------------------------
#Pruning the ANN
#For this best fitting model, look at the smallest magnitude for the different weights and drop/set these to 0

#Seed to replicate results
set.seed(23456)
#Fit neural network model
neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
                     linout = FALSE, decay = 0.01, entropy = TRUE, maxit = 500)

#Summarize structure of neural network model
summary(neuralnetmod)

#Set the weights equal to false 'F' or true 'T' depending on whether you are pruning the weight connection.
pick.weights <- as.logical(c(F, F, T, F, F, F, T, F, T, F,
                             T, F, F, T, F, F, F, F, F, F,
                             T, T))

#Set the chosen weights equal to zero corresponding to the above order
start.weights <- neuralnetmod$wts
start.weights[c(8,4,16,17,19,13,2,15,18,10,6,1,12,5,20)] <- 0

#Set seed to replicate results
set.seed(23456)
#Run the pruned network
neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
                     linout = FALSE, decay = 0.01, entropy = TRUE, maxit = 500, Wts = start.weights, mask = pick.weights)

#Summarize structure of neural network model
summary(neuralnetmod)
#Rerun cross validation function as above to obtain
#cross validated deviance

# ------------------------------------------------------------------------------------------------------------------
#The best fitting model has 18 weights

#Seed to replicate results
set.seed(23456)
#Fit neural network model
neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
                     linout = FALSE, decay = 0.01, entropy = TRUE, maxit = 500)
#Summarize structure of neural network model
summary(neuralnetmod)
```

```
#Set the 22 weights equal to false 'F' or true 'T' depending on whether you are pruning the weight connection.
pick.weights <- as.logical(c(T, T, T, F, T, T, T, F, T, T,
                             T, T, T, T, T, F, F, T, T, T,
                             T, T))

#Set the chosen weights equal to zero corresponding to the above order
start.weights <- neuralnetmod$wts
start.weights[c(8,4,16,17)] <- 0

#Set seed to replicate results
set.seed(23456)
#Run the pruned network
neuralnetmod <- nnet(Binary.outcome~stan.FIT + stan.age + GENDER + prev.incident, data=ccFIT20, skip = FALSE, size = 3,
                     linout = FALSE, decay = 0.01, entropy = TRUE, maxit = 500, Wts = start.weights, mask = pick.weights)



#Summarize structure of neural network model
summary(neuralnetmod)
#Rerun cross validation function as above to obtain
#cross validated deviance
# ------------------------------------------------------------------------------------------------------------------
#Plot to show change in CV deviance as weight connections removed

pruningcvdeviance <- read.csv("pruningcvdeviance.csv")
pruningcvdeviance$Number.of.Weights<-as.numeric(pruningcvdeviance$Number.of.Weights)
attach(pruningcvdeviance)

dev.new(width=10, height=10)
plot(Number.of.Weights,Cross.Validated.Deviance, xlab="Number of Weights", ylab="Cross-Validated Deviance", xlim=c(22,
6))
lines(Number.of.Weights,Cross.Validated.Deviance,col="black")
axis(1, at=c(22:7), labels= c(22:7))

# ------------------------------------------------------------------------------------------------------------------
#Garsons algorithm to visualise relative importance of variables (for final selected model)

cols <- blues9
garson(neuralnetmod) +
  scale_y_continuous('Relative Importance', limits = c(-1, 1)) +
  scale_fill_gradientn(colours = cols) +
  scale_colour_gradientn(colours = "black") +
  theme(text = element_text(size=15),
        axis.text.x = element_text(angle=45, hjust=1)) +
  scale_x_discrete(labels=c("Sex","FIT Result","Age","Previous Non-Responder","Previous Responder")) +
  xlab(NULL)

# ------------------------------------------------------------------------------------------------------------------
#Performance Measures:
# ------------------------------------------------------------------------------------------------------------------
#Calibration plot

install.packages("PredictABEL")
library("PredictABEL")
# ------------------------------------------------------------------------------------------------------------------
#Calibration Plot (for the neural network model)

cOutcome <- 13 #the column with the outcome in your dataset
predRisk <- fitted.values(neuralnetmod)
# specify range of x-axis and y-axis
rangeaxis <- c(0,1)
# specify number of groups for Hosmer-Lemeshow test
groups <- 10

cal <- plotCalibration(data=ccFIT20, cOutcome=cOutcome, predRisk=predRisk, groups=groups, rangeaxis=rangeaxis)

#Return observed versus expected probability table and p value for Hosmer-Lemeshow goodness of fit test
cal

#------------------------------------------------------------------------------------------------------------------
```

```
#Another way to determine Hosmer Lemeshow statistic

install.packages("ResourceSelection")
library("ResourceSelection")

hl <-hoslem.test(ccFIT20$Binary.outcome, fitted.values(neuralnetmod), g=10)
hl

#loop to give results for different group splits

for (i in 5:15) {
  print(hoslem.test(ccFIT20$Binary.outcome, fitted.values(neuralnetmod), g=i)$p.value)
}

# ------------------------------------------------------------------------------------------------------------
#Predictiveness curve for FIT only, logistic regression and ANN
# ------------------------------------------------------------------------------------------------------------

#FIT only model
cclog.mod.1 <- glm(Binary.outcome~ log(TRANS_FIT_KIT_RESULT+1), ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.1)

#FIT plus risk model (logistic regression)
cclog.mod.2 <- glm(Binary.outcome ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident,
data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.2)

#Neural Network
neuralnetmod
summary(neuralnetmod)

#Different methods for predicted probabilities

#predicted probabilities for neural network
ccFIT20$nnetprobabilities <-fitted.values(neuralnetmod)
ccFIT20$nnetprobabilities <- as.vector(ccFIT20$nnetprobabilities)

#FIT only
ccFIT20$fitprobabilities <- predict(cclog.mod.1, ccFIT20, type = "response")

#Predicted probabilities for just the logistic regression model
ccFIT20$lrprobabilities <- predict(cclog.mod.2, ccFIT20, type = "response")

#Other method:

predRisk1 <- predRisk(cclog.mod.1)
predRisk2 <- predRisk(cclog.mod.2)
predRisk3 <- fitted.values(neuralnetmod)
# ------------------------------------------------------------------------------------------------------------
# Predictiveness curve

# specify range of y-axis
rangeyaxis <- c(0,1)
# specify labels of the predictiveness curves
labels <- c("Neural Network", "Risk adjusted LR", "FIT only")


# produce predictiveness curves
plotPredictivenessCurve(predrisk=cbind(predRisk3,predRisk2, predRisk1),rangeyaxis=rangeyaxis, labels=labels)

# ------------------------------------------------------------------------------------------------------------
# ------------------------------------------------------------------------------------------------------------
#Plotting ROC curve for all three models

#Install pROC packages
install.packages("pROC")
library("pROC")

# ------------------------------------------------------------------------------------------------------------
```

```
# ------------------------------------------------------------------------------------------------------------
#1st ROC for FIT only

#1st ROC curve for FIT only
roccurve1 <- roc(ccFIT20$Binary.outcome ~ ccFIT20$fitprobabilities)
#Return AUC
roccurve1

#Alternatively for 95% CI:
auc(ccFIT20$Binary.outcome, ccFIT20$fitprobabilities)
ci.auc(ccFIT20$Binary.outcome, ccFIT20$fitprobabilities, conf.level=0.95, method="delong")

#Return AUC CI
ci.auc(roccurve1, conf.level=0.95, method="delong")
ci.auc(roccurve1, conf.level=0.95, method="bootstrap", boot.n = 10000)

# ------------------------------------------------------------------------------------------------------------
#2nd ROC curve for Risk adjusted

roccurve2 <- roc(ccFIT20$Binary.outcome ~ ccFIT20$lrprobabilities)

#Return AUC
roccurve2

#Return AUC CI
ci.auc(roccurve2, conf.level=0.95, method="delong")
ci.auc(roccurve2, conf.level=0.95, method="bootstrap", boot.n = 10000)
# ------------------------------------------------------------------------------------------------------------
#3rd ROC curve for Neural Network

roccurve3 <- roc(ccFIT20$Binary.outcome ~ ccFIT20$nnetprobabilities)

#Return AUC
roccurve3

#Return AUC CI
ci.auc(roccurve3, conf.level=0.95, method="delong")
ci.auc(roccurve3, conf.level=0.95, method="bootstrap", boot.n = 10000)

# ------------------------------------------------------------------------------------------------------------

#Combine all three models in a ROC plot

#ROC for risk adjusted model
Risk_adjusted_LR <- plot.roc(ccFIT20$Binary.outcome, ccFIT20$lrprobabilities)

#ROC for neural network
Neural_Network <- plot.roc(ccFIT20$Binary.outcome, ccFIT20$nnetprobabilities, add=TRUE, col="12", lty=2)

#Add ROC for FIT only
FIT_only <- plot.roc(ccFIT20$Binary.outcome, ccFIT20$fitprobabilities, add=TRUE, col="red", lty=3)

#Add legend
legend("right", legend = c("Risk-adjusted LR", "Neural Network", "FIT only"), lty=c(1, 2, 3), col=c("black", "12", "red"))

# ------------------------------------------------------------------------------------------------------------

#ROC test - to test for significant difference between neural network and logistic regression

roc.test(roccurve2, roccurve3)

# The latter used Delong's test. To use bootstrap test:
#Set seed for replicability
roc.test(roccurve2, roccurve3, method="bootstrap", boot.n=10000)
```

```
# ------------------------------------------------------------------------------------------------------------------------
#Test performance 2 by 2 data
# ------------------------------------------------------------------------------------------------------------------------
#Install dplyr for data manipulation

install.packages("dplyr")
library("dplyr")

# ------------------------------------------------------------------------------------------------------------------------
#Produce predicted probabilities for indivduals using all three models
#FIT only model
cclog.mod.1 <- glm(Binary.outcome~ log(TRANS_FIT_KIT_RESULT+1), ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.1)

#FIT plus risk model (logistic regression)
cclog.mod.2 <- glm(Binary.outcome ~ log(TRANS_FIT_KIT_RESULT + 1) + AGE_AT_EPISODE_START + GENDER + prev.incident,
data = ccFIT20, family=binomial(link="logit"))
summary(cclog.mod.2)

#Neural Network
neuralnetmod
summary(neuralnetmod)

#Different methods for predicted probabilities

#predicted probabilities for neural network
ccFIT20$nnetprobabilities <-fitted.values(neuralnetmod)
#To transform into numerical vector
ccFIT20$nnetprobabilities<- as.vector(ccFIT20$nnetprobabilities)
#Alternative
ccFIT20$nnetprobabilities <-as.numeric(fitted.values(neuralnetmod))
str(ccFIT20$nnetprobabilities)

#FIT only
ccFIT20$fitprobabilities <- predict(cclog.mod.1, ccFIT20, type = "response")

#Predicted probabilities for just the logistic regression model
ccFIT20$lrprobabilities <- predict(cclog.mod.2, ccFIT20, type = "response")

#To give full decimal places create character variable for predicted probabilities

ccFIT20$CHARnnetprob <- as.character(ccFIT20$nnetprobabilities)
ccFIT20$CHARlrprob <- as.character(ccFIT20$fitprobabilities)
ccFIT20$CHARfitprob <- as.character(ccFIT20$lrprobabilities)

# ------------------------------------------------------------------------------------------------------------------------
# Set up a 160 cutpoint for the FIT

ccFIT20$cutpoint160[ccFIT20$TRANS_FIT_KIT_RESULT>=160] <- "ABNORMAL"
ccFIT20$cutpoint160[ccFIT20$TRANS_FIT_KIT_RESULT<160] <- "NORMAL"

# Convert the column to a factor variable
ccFIT20$cutpoint160 <- factor(ccFIT20$cutpoint160)

#Look at the numbers referred for this cutpoint and set the same
xtabs(~ccFIT20$cutpoint160+ccFIT20$Binary.outcome)

#375 referred
ccFIT20[375,]

#Arrange by predicted probabilities for the model and refer the same

ccFIT20<- arrange(ccFIT20, desc(nnetprobabilities))

#Determine the neural network probability cutpoint
ccFIT20[375,]

#Set this risk probability for a positive or negative test result
ccFIT20$riskthreshold160[ccFIT20$nnetprobabilities<"0.40719445665783"] <- "NORMAL"
ccFIT20$riskthreshold160[ccFIT20$nnetprobabilities>="0.40719445665783"] <- "ABNORMAL"
```

```
#160 risk cutpoint 2 by 2 table

xtabs(~ccFIT20$riskthreshold160+ccFIT20$Binary.outcome)

#160 risk cutpoint 2 by 2 table for sex

xtabs(~ccFIT20$riskthreshold160+ccFIT20$Binary.outcome + ccFIT20$GENDER)

#160 risk cutpoint 2 by 2 table for severity and sex

xtabs(~ccFIT20$riskthreshold160+ccFIT20$Binary.outcome + ccFIT20$GENDER + ccFIT20$OUTCOME)

# ------------------------------------------------------------------------------------------------------------
#Repeat this process for cutpoints from 30 to 180 and for all three models to populate two by two table
# ------------------------------------------------------------------------------------------------------------
#e.g. for logistic regression model

#Arrange by predicted probabilities for the model and refer the same

ccFIT20<- arrange(ccFIT20, desc(lrprobabilities))

#375 referred
ccFIT20[375,]

#Assign a 160 risk probability cutpoint

ccFIT20$riskthreshold160[ccFIT20$lrprobabilities<"0.389255384060317"] <- "NORMAL"
ccFIT20$riskthreshold160[ccFIT20$lrprobabilities>="0.389255384060317"] <- "ABNORMAL"

#160 risk cutpoint 2 by 2 tables

xtabs(~ccFIT20$riskthreshold160+ccFIT20$Binary.outcome)
xtabs(~ccFIT20$riskthreshold160+ccFIT20$Binary.outcome+ccFIT20$GENDER+ccFIT20$OUTCOME)


# ------------------------------------------------------------------------------------------------------------
```

Appendix 2: Weight Connection Values for a 5-3-1 Neural Network

Node	Weight Connections
b->h1	-2.76
i1->h1	-0.11
i2->h1	-0.87
i3->h1	0.01
i4->h1	1.95
i5->h1	1.93
b->h2	-2.94
i1->h2	-0.05
i2->h2	-0.53
i3->h2	0.03
i4->h2	1.24
i5->h2	1.21
b->h3	-33.84
i1->h3	-36.30
i2->h3	3.41
i3->h3	19.15
i4->h3	-48.50
i5->h3	-25.82
b->o	-6.25
h1->o	-69.04
h2->o	171.40
h3->o	-1.64

Table A.2.1: A feed forward 5-3-1 neural network with 22 weights, 500 iterations and 0 weight decay. Neural network plotted using nnet and the neuralnetworktools packages in R.

Appendix 3: Hosmer-Lemeshow goodness of fit test for different splits for the ANN

Number of groups	P value for the Hosmer and Lemeshow test (ANN)	P value for the Hosmer and Lemeshow test (Logistic Regression)
5	0.993	0.906
6	0.974	0.716
7	0.746	0.611
8	0.955	0.802
9	0.969	0.793
10	0.892	0.898
11	0.978	0.647
12	0.943	0.806
13	0.986	0.989
14	0.959	0.798
15	0.912	0.940

Table A.3.1: Hosmer-Lemeshow goodness of fit test for different splits for the ANN

Refined ANN Calibration	Total	Mean predicted probability	Mean observed probability	Predicted (n)	Observed (n)
[0.0634,0.146)	181	0.114	0.099	20.56	18
[0.1464,0.181)	181	0.164	0.171	29.75	31
[0.1809,0.215)	181	0.199	0.199	36.11	36
[0.2149,0.248)	181	0.233	0.232	42.10	42
[0.2476,0.276)	181	0.262	0.260	47.41	47
[0.2760,0.315)	181	0.295	0.304	53.37	55
[0.3148,0.359)	181	0.339	0.298	61.34	54
[0.3588,0.411)	181	0.384	0.431	69.45	78
[0.4115,0.503)	181	0.456	0.448	82.45	81
[0.5030,0.820]	181	0.588	0.591	106.49	107

Chi square – 3.586 df – 8 p value 0.8924

Table A.3.2: Hosmer lemeshow statistics for the refined neural network

Appendix 4: Two by two tables for FIT only, Risk-adjusted and Neural Network models at thresholds between 30-180 µg/g

30 µg Hb/g faeces Threshold	Diagnostic Positive						Diagnostic Negative						Total
	FIT only		Risk-adjusted		Neural Network		FIT only		Risk-adjusted		Neural Network		
	Male	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
FIT/Risk Positive	***Total***	**324**	**161**	**357**	**139**	**353**	**152**	***Total***	**500**	**481**	**615**	**355**	**586**	**375**	
	Cancer	48	21	50	21	50	21	***Low Risk Adenoma***	212	161	257	123	247	127	
	High risk Adenoma	127	56	145	55	144	57	***Abnormal***	198	208	245	152	232	166	**1466**
	Intermediate risk Adenoma	149	84	162	63	159	74	***Normal (No Abnormalities Found)***	90	112	113	80	107	82	
FIT/Risk Negative		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
	Total	**34**	**30**	**1**	**52**	**5**	**39**	***Total***	**132**	**148**	**17**	**274**	**46**	**254**	
	Cancer	2	2	0	2	0	2	***Low Risk Adenoma***	51	42	6	80	16	76	
	High risk Adenoma	18	13	0	14	1	12	***Abnormal***	51	74	4	130	17	116	**344**
	Intermediate risk Adenoma	14	15	1	36	4	25	***Normal (No Abnormalities Found)***	30	32	7	64	13	62	
Total	**549**						**1261**								**1810**

Table A.4.1: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 30 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.191 for the risk-adjusted model and 0.178 for the neural network. Profiles of outcome severity are also given.

40 µg Hb/g faeces Threshold	Diagnostic Positive						Diagnostic Negative						Total
	FIT only		Risk-adjusted		Neural Network		FIT only		Risk-adjusted		Neural Network		
	Male	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
FIT/Risk Positive	***Total***	**282**	**139**	**347**	**93**	**343**	**103**	***Total***	**406**	**364**	**568**	**183**	**543**	**202**	
	Cancer	45	20	49	19	48	18	***Low Risk Adenoma***	173	120	237	65	229	65	
	High risk Adenoma	115	51	144	37	141	42	***Abnormal***	157	157	227	78	214	90	**1191**
	Intermediate risk Adenoma	122	68	154	37	154	43	***Normal (No Abnormalities Found)***	76	87	104	40	100	47	
FIT/Risk Negative		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
	Total	**76**	**52**	**11**	**98**	**15**	**88**	***Total***	**226**	**265**	**64**	**446**	**89**	**427**	
	Cancer	5	3	1	4	2	5	***Low Risk Adenoma***	90	83	26	138	34	138	
	High risk Adenoma	30	18	1	32	4	27	***Abnormal***	92	125	22	204	35	192	**619**
	Intermediate risk Adenoma	41	31	9	62	9	56	***Normal (No Abnormalities Found)***	44	57	16	104	20	97	
Total	**549**						**1261**								**1810**

Table A.4.2 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 40 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.242 for the risk-adjusted model and 0.232 for the neural network. Profiles of outcome severity are also given.

50 µg Hb/g faeces Threshold	Diagnostic Positive						Diagnostic Negative						Total
	FIT only		Risk-adjusted		Neural Network		FIT only		Risk-adjusted		Neural Network		
	Male	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
FIT/Risk Positive	***Total***	**254**	**125**	**323**	**66**	**307**	**94**	***Total***	**342**	**288**	**489**	**131**	**455**	**153**	
	Cancer	40	19	46	12	42	17	***Low Risk Adenoma***	145	97	208	48	194	48	
	High risk Adenoma	104	45	137	25	132	38	***Abnormal***	140	120	190	52	182	65	**1009**
	Intermediate risk Adenoma	110	61	140	29	133	39	***Normal (No Abnormalities Found)***	57	71	91	31	79	40	
FIT/Risk Negative		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
	Total	**104**	**66**	**35**	**125**	**51**	**97**	***Total***	**290**	**341**	**143**	**498**	**177**	**476**	
	Cancer	10	4	4	11	8	6	***Low Risk Adenoma***	118	106	55	155	69	155	
	High risk Adenoma	41	24	8	44	13	31	***Abnormal***	109	162	59	230	67	217	**801**
	Intermediate risk Adenoma	53	38	23	70	30	60	***Normal (No Abnormalities Found)***	63	73	29	113	41	104	
Total	**549**						**1261**								**1810**

Table A.4.3: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 50 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.272 for the risk-adjusted model and 0.260 for the neural network. Profiles of outcome severity are also given.

80 µg Hb/g faeces Threshold	Diagnostic Positive							Diagnostic Negative							Total
		FIT only		Risk-adjusted		Neural Network			FIT only		Risk-adjusted		Neural Network		
		Male	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
FIT/Risk Positive	***Total***	**190**	**91**	**248**	**41**	**233**	**71**	***Total***	**198**	**189**	**306**	**73**	**273**	**92**	**668** (669 for neuralnet)
	Cancer	36	17	37	11	38	15	***Low Risk Adenoma***	83	65	125	24	107	30	
	High risk Adenoma	72	35	112	17	102	33	***Abnormal***	81	79	121	32	117	37	
	Intermediate risk Adenoma	82	39	99	13	93	23	***Normal (No Abnormalities Found)***	34	45	60	17	49	25	
FIT/Risk Negative		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	**1142** (1141 for neuralnet)
	Total	**168**	**100**	**110**	**150**	**125**	**120**	***Total***	**434**	**440**	**326**	**556**	**359**	**537**	
	Cancer	14	6	13	12	12	8	***Low Risk Adenoma***	180	138	138	179	156	173	
	High risk Adenoma	73	34	33	52	43	36	***Abnormal***	168	203	128	250	132	245	
	Intermediate risk Adenoma	81	60	64	86	70	76	***Normal (No Abnormalities Found)***	86	99	60	127	71	119	
Total				**549**							**1261**				**1810**

Table A.4.4: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 80 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.321 for the risk-adjusted model and 0.330 for the neural network. Profiles of outcome severity are also given.

150 µg Hb/g faeces Threshold	Diagnostic Positive						Diagnostic Negative						Total
	FIT only		Risk-adjusted		Neural Network		FIT only		Risk-adjusted		Neural Network		
	Male	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
FIT/Risk Positive	***Total***	**122**	**56**	**163**	**30**	**151**	**56**	***Total***	**118**	**104**	**172**	**35**	**138**	**55**	
	Cancer	29	11	30	9	28	10	***Low Risk Adenoma***	46	34	67	9	51	15	
	High risk Adenoma	48	21	75	13	72	28	***Abnormal***	53	44	70	15	60	22	**400**
	Intermediate risk Adenoma	45	24	58	8	51	18	***Normal (No Abnormalities Found)***	19	26	35	11	27	18	
FIT/Risk Negative		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
	Total	**236**	**135**	**195**	**161**	**207**	**135**	***Total***	**514**	**525**	**460**	**594**	**494**	**574**	
	Cancer	21	12	20	14	22	13	***Low Risk Adenoma***	217	169	196	194	212	188	
	High risk Adenoma	97	48	70	56	73	41	***Abnormal***	196	238	179	267	189	260	**1410**
	Intermediate risk Adenoma	118	75	105	91	112	81	***Normal (No Abnormalities Found)***	101	118	85	133	93	126	
Total	**549**						**1261**								**1810**

Table A.4.5: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 150 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.383 for the risk-adjusted model and 0.399 for the neural network. Profiles of outcome severity are also given.

170 µg Hb/g faeces Threshold	Diagnostic Positive							Diagnostic Negative							Total
		FIT only		Risk-adjusted		Neural Network			FIT only		Risk-adjusted		Neural Network		
		Male	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
FIT/Risk Positive	***Total***	**112**	**52**	**149**	**25**	**145**	**43**	***Total***	**105**	**93**	**155**	**33**	**128**	**46**	
	Cancer	27	10	29	7	27	10	***Low Risk Adenoma***	37	27	57	9	49	13	
	High risk Adenoma	42	20	69	11	71	23	***Abnormal***	50	40	66	14	54	18	**362**
	Intermediate risk Adenoma	43	22	51	7	47	10	***Normal (No Abnormalities Found)***	18	26	32	10	25	15	
FIT/Risk Negative		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*		*Male*	*Female*	*Male*	*Female*	*Male*	*Female*	
	Total	**246**	**139**	**209**	**166**	**213**	**148**	***Total***	**527**	**536**	**477**	**596**	**504**	**583**	
	Cancer	23	13	21	16	23	13	***Low Risk Adenoma***	226	176	206	194	214	190	
	High risk Adenoma	103	49	76	58	74	46	***Abnormal***	199	242	183	268	195	264	**1448**
	Intermediate risk Adenoma	120	77	112	92	116	89	***Normal (No Abnormalities Found)***	102	118	88	134	95	129	
Total				**549**							**1261**				**1810**

Table A.4.6: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 170 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.392 for the risk-adjusted model and 0.411 for the neural network. Profiles of outcome severity are also given.

180 µg Hb/g faeces Threshold	Diagnostic Positive						Diagnostic Negative						Total
	FIT only		Risk-adjusted		Neural Network		FIT only		Risk-adjusted		Neural Network		
	Male	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
FIT/Risk Positive	***Total***	**108**	**49**	**142**	**23**	**142**	**38**	***Total***	**97**	**85**	**145**	**29**	**121**	**38**	
	Cancer	26	10	29	6	25	9	***Low Risk Adenoma***	37	25	53	7	49	9	
	High risk Adenoma	41	17	65	11	71	21	***Abnormal***	42	37	63	13	50	15	**339**
	Intermediate risk Adenoma	41	22	48	6	46	8	***Normal (No Abnormalities Found)***	18	23	29	9	22	14	
FIT/Risk Negative		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***		***Male***	***Female***	***Male***	***Female***	***Male***	***Female***	
	Total	**250**	**142**	**216**	**168**	**216**	**153**	***Total***	**535**	**544**	**487**	**600**	**511**	**591**	
	Cancer	24	13	21	17	25	14	***Low Risk Adenoma***	226	178	210	196	214	194	
	High risk Adenoma	104	52	80	58	74	48	***Abnormal***	207	245	186	269	199	267	**1471**
	Intermediate risk Adenoma	122	77	115	93	117	91	***Normal (No Abnormalities Found)***	102	121	91	135	98	130	
Total	**549**						**1261**								**1810**

Table A.4.7: 2 by 2 table for FIT only, the risk-adjusted logistic regression model and the neural network split by sex. A threshold of 180 µg Hb/g faeces was used for the FIT which is equivalent to a risk threshold of 0.399 for the risk-adjusted model and 0.425 for the neural network. Profiles of outcome severity are also given.

Appendix 5: Cancer/advanced adenoma detection rate for each model by screening history and sex subgroup

Cancer/advanced adenoma detection rate for each model by screening history and sex subgroup (Threshold 160 µg Hb/g)															
	FIT Only					Risk Model					Neural Network				
Subgroup	**TP**	**FP**	**FN**	**TN**	**Cancer/Advanced Adenoma Detection Rate (%)**	**TP**	**FP**	**FN**	**TN**	**Cancer/Advanced Adenoma Detection Rate (%)**	**TP**	**FP**	**FN**	**TN**	**Cancer/Advanced Adenoma Detection Rate (%)**
Female First Time Invitee	4	10	12	64	4.44	0	1	16	73	0.00	0	0	16	74	0.00
Male First Time Invitee	12	14	13	61	12.00	5	6	20	69	5.00	4	6	21	69	4.00
Female Non Responder	14	10	18	49	15.38	14	12	18	47	15.38	16	7	16	52	17.58
Male Non Responder	27	15	47	70	16.98	59	63	15	22	37.11	47	29	27	56	29.56
Female Responder	36	76	107	420	5.63	12	22	131	474	1.88	31	42	112	454	4.85
Male Responder	76	81	183	391	10.40	92	89	167	383	12.59	95	98	164	374	13.00
TP – True Positive; FP – False Positive; FN- False Negative; TN – True Negative															

Table A.5.1: Cancer/advanced adenoma detection rate for each model by screening history and sex subgroup using a threshold of 160 µg Hb/g for the FIT and equivalent risk thresholds for the ANN model and logistic regression model.

Investigating the Use of Routine GP Patient Data to Improve Colorectal Cancer Screening Referral Decisions

This research was carried out as part of an NIHR Infrastructure Doctoral Training Exchange (IDTE) Award based at the Institute of Applied Health Research at The University of Birmingham. IDTE supervisors were Professor Tom Marshall (TM) and Dr Ronan Ryan (RR) based in the Health Informatics team, Primary Care Division.

ABSTRACT

Background: Risk prediction models which incorporate faecal occult blood tests with other known colorectal cancer risk factors have demonstrated increased sensitivity compared with FOBT alone. One of the barriers for using these prediction models in practice is the collection of data from individuals. Databases of electronic patient records from primary care have a richer level of data than that available on the bowel cancer screening system (BCSS) and include details on, symptoms, diagnoses, prescriptions, laboratory test results, socioeconomic status, lifestyle parameters and anthropometrics. These clinical features may add a further dimension to a risk based prediction model to improve colorectal cancer screening referral decisions. Further information from GP databases could be drawn out onto the screening system to contribute to decision making. The aim of the study was to determine (i) the availability of GP data for key predictors of colorectal cancer in the screening population and (ii) whether this additional information has the potential to be used to make more accurate screening referral decisions by developing multivariable risk prediction models.

Methods: The Health Improvement Network (THIN) database of anonymised GP records was used to define a screening population by identifying practices which receive electronic bowel cancer screening programme notifications in England and for participants aged 60-74. The positivity (number of positive results out of all FOBT results) of the FOBT in this cohort was compared to that in the literature. The availability of predictors was investigated by determining the percentage recorded on the database for each variable. Univariable analysis using Cox Regression was used to estimate hazard ratios (HR) for >30 key clinical features of colorectal cancer driven from the literature. Kaplan-Meier estimates for time to diagnosis and time to death were plotted for a population with both negative and positive FOBTs. These were stratified by test result type and by sex with a log rank test determining if there was a statistical difference between these groups. A risk prediction

model combining the FOBT with risk predictors was then developed using Cox Regression with backwards elimination (p value >0.05 as removal criterion). Multivariable fractional polynomial models for continuous variables and interactions were investigated. To determine model performance, Harrell's C statistic and Somers' D were determined, along with plots of the Cox-Snell residuals to determine goodness of fit. Cox Regression diagnostics were also investigated by looking at Schoenfeld residual plots and tests of proportionality for individual covariates. The baseline hazard was estimated at 2 years using a non-parametric survival estimate (Kaplan-Meier). Absolute risk predictions for individuals were then estimated. Parametric survival models were investigated to determine whether these provided a better fit to the data using the AIC cumulative hazard, Kaplan Meier function plots and Cox-Snell residuals. Since the guaiac FOBT sensitivity is around 50%, analysis was then repeated for a population with negative FOBT results only to determine whether additional factors could be used for screening referral decisions despite a negative test result. A Nomogram was produced for this model as an alternative method of applying and presenting the risk prediction model.

Results: The screening cohort derived from THIN gave 292,168 patients across 360 practices aged 60-74 with a positive or negative FOBT result. The most severe diagnosis within 2 years was colorectal cancer for 929 patients and polyps for 1960 patients (2889 total). Data were generally well recorded, with binary variables (like symptoms) having 100% recorded completeness, smoking status 99.44% completeness and alcohol consumption 78% completeness. The least complete factors included lab results, e.g. platelet count, mean cell volume, and haemoglobin at around 45%, and ferritin at 8.59%. Univariable Cox Regression identified that screening based factors had the strongest association with colorectal cancer/polyps. For example, previous positive FOBT results had a HR of 5.032 (CI: 4.184-6.052) and previous polyps diagnosed before the latest FOBT result had a HR of 3.182 (CI: 2.768-3.659). The symptom with the highest HR was rectal bleeding (HR 3.118 (2.503-3.883)). A Cox regression model was developed for those with a positive/negative screening test (n=98,303, 1197 events). The final model included: FOBT result, smoking status, whether a patient has a diagnosis of Crohn's disease, previous polyps diagnosed, flatulence, MCV of <80fL compared to a MCV of ≥80fL, alcohol consumption in units per week, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit. Significant interactions at the 0.05 p-value level included FOBT result and age and MCV and age. The optimism adjusted performance metrics gave a; C statistic of 0.850, c-slope of

0.991, D statistic 2.298 and R^2 of 0.558. The best fitting parametric model based on the AIC, cumulative hazard plots, Kaplan Meier function plots and Cox-Snell residuals was the generalised gamma model which uses the accelerated failure time metric. A model investigating predicted probabilities for those with just negative results was also developed to explore whether factors could be used to determine whether a patient should be referred despite a negative result. The final model for those with a negative FOBT only (n = 95,792, 587 events) contained the following variables: smoking status, whether a patient has an irritable bowel syndrome (IBS) diagnosis, previous polyps diagnosed, flatulence, weight loss, MCV of <80fL compared to a MCV of ≥80fL, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit. Optimism adjusted performance metrics for this model gave; a C statistic of 0.650, c-slope of 0.944, D statistic 0.836 and R^2 of 0.144. A nomogram was produced for this model as a visual representation. The best fitting parametric model for negative results only was the Gompertz model.

Conclusions: This chapter has shown that there are several clinical predictors available from GP databases which are associated with colorectal cancer and polyps for an English screening population. Furthermore, this research has identified predictors which could be considered for inclusion in a future risk adjusted screening model. Most factors contributing to the risk based models are well recorded. Laboratory parameters, although shown to be associated with colorectal cancer diagnosis, are the least well recorded factors from the model. Additional potential predictors which could be considered for inclusion in a future risk adjusted model include those which relate to screening history which have a strong association with the diagnosis of colorectal cancer/polyps. Predictors which retained significance in both multivariable models included demographic factors age and sex, lifestyle factor smoking status, lab test factor MCV, family history of gastrointestinal cancer, previous polyps diagnosed and the following symptoms: abdominal pain/antispasmodic prescription, diarrhoea, flatulence, and change in bowel habit. Additional data could potentially be drawn from primary care onto the BCSS if these factors are shown to contribute to the assessment of an individual's risk of colorectal cancer. Similar analyses could be carried out with the FIT which is due to be introduced to the English NHS BCSP. Although, there would be many issues relating to quality of data, how the data is handled by the GP and the use of different GP operating systems in the NHS. Future risk screening

models including the identified factors could be used to help make more accurate screening referral decisions to improve early detection of colorectal cancer.

1.0 INTRODUCTION

The previous chapters have investigated the use of routine screening data from the Bowel Cancer Screening System (BCSS) along with the Faecal Immunochemical Test (FIT) to develop prediction models and aid referral decisions for colorectal cancer based on risk. **Chapter 3** used standard statistical methodology in the form of logistic regression and **Chapter 4** extended this to include a machine learning approach in the form of neural networks. Although there was a statistically significant improvement in discrimination when applying the neural network model, the value of the AUC ROC of the models suggest that there is room for improvement in prediction which could be achieved through other risk information/participant data. **Chapter 2** identified that lifestyle factors and lab results could enhance the discriminatory power in some of the included studies.[1] Other sources of data and individual participant information could therefore be used to add a further dimension to a risk based score and improve model performance and test accuracy.

As health care records are becoming increasingly electronic and more quality assurance processes are being implemented to enhance data completeness and accuracy, databases of electronic patient records from general practices offer an additional way to make use of routinely available information on individuals. GP records have a richer level of data than that available from the BCSS and include details on symptoms, diagnoses, prescriptions, laboratory test results, socioeconomic status, lifestyle parameters and anthropometrics.

This chapter investigated the use of an anonymized GP record database (THIN – The Health Improvement Network) to determine how complete potential predictors are for this population and identified predictors which could be considered for inclusion in a future risk adjusted screening model.

1.1 Primary Care Databases for Research

There are a number of large primary care databases available for medical research in the UK; primary amongst these are the Clinical Practice Research Datalink (CRPD) (formerly General Practice Research Database - GPRD), The Health Improvement Network (THIN), QRESEARCH and the IMS Mediplus system.[2]

The THIN database of anonymised GP records has data for over 587 practices (5.67% coverage of UK in 2014) covering more than 12 million patients (including 3.6 million active patients).[3] THIN provides information on diagnoses, symptoms, prescriptions, laboratory tests and lifestyle factors across four standardised data files.[4] These include patient files (recording information such as age, sex and other demographics), medical records (diagnoses and symptoms), therapy records (prescriptions) and additional health data (AHD) records (smoking, cholesterol etc). These different file types are linked by a patient ID.

Diagnoses and symptoms are currently recorded as hierarchical Read codes and prescriptions are linked to the British National Formulary (BNF) chapter codes. Read codes are clinical codes which the GP enters for a patient consultation relating to new diagnoses and symptoms.

1.2 Read Codes from the Bowel Cancer Screening Programme used in Primary Care

Relevant Read codes generated for the English BCSP and used as electronic notifications to primary care are listed in **Table 1**. Read codes are used on primary care systems to record various symptoms and diagnoses and can be drawn out of databases for analysis.

Read Code	Description
6866	Bowel cancer screening programme: faecal occult blood result
6867	Bowel cancer screening programme faecal occult blood testing kit spoilt
686A	Bowel cancer screening programme faecal occult blood test normal
686B	Bowel cancer screening programme faecal occult blood test abnormal
686C	Bowel cancer screening programme faecal occult blood testing incomplete participation
9Ow2	No response to bowel cancer screening programme invitation

Table 1: Set of Read codes used by the NHS BCSP to record colorectal cancer screening activity.

Currently there are no specific Read codes for the FIT, which is likely to be implemented in the BCSP in Summer 2018. The current screening test for bowel cancer is the guaiac based FOBT (Hemascreen, Immunostics, New Jersey, USA) which identifies individuals at increased risk of cancer by detecting blood from colorectal neoplasia.[5] A systematic review suggests that the FIT has a sensitivity of 87.2% and a specificity of 92.8% whereas the gFOBT has a sensitivity of 47.4% and a specificity of 92%.[6] The FIT however is currently not routinely used for screening in the UK and current GP records only receive electronic notification results for the gFOBT.

Health systems in the UK are also transitioning over to SNOMED CT clinical terminology and so Read codes will eventually cease to be used. As the FIT is implemented in the UK, GP

records will begin to obtain more data on the FIT and the approaches used in this chapter could be used to investigate prediction models combining this newer test.

1.3 Links between GP records and the Bowel Cancer Screening system

When patients sign up to a GP practice, their details are uploaded to the NHS Information Authority.[7] The Spine supports IT infrastructure for health and social care in England and is a set of services used by the NHS Care Record Service including the Personal Demographics Service (demographic information and NHS number of patients) and Summary Care Record (clinical information of a patient).[8] The Spine draws out these registration details over-night. [8] Correspondingly the details of everyone who falls within the age range of screening (60-74) is extracted to the BCSS to gain new patient details. The Spine is set to what information is drawn from the GP practice and includes name, address and birth date. There is scope to draw out additional information from primary care such as laboratory test results, other co-morbidities/conditions and lifestyle parameters. A schematic of how the BCSS links up with the other data services used within the NHS and Public Health England is shown in **Figure 1**.

Since 2010, The BCSS has sent results of the FOBT electronically to GP practices who have opted for this service using the same system as the Pathology Messaging Implementation Programme (PMIP). [9] This is the service also used by pathology labs to send laboratory results to primary care. The gFOBT results from screening are sent once daily overnight to the GP practices in batches.

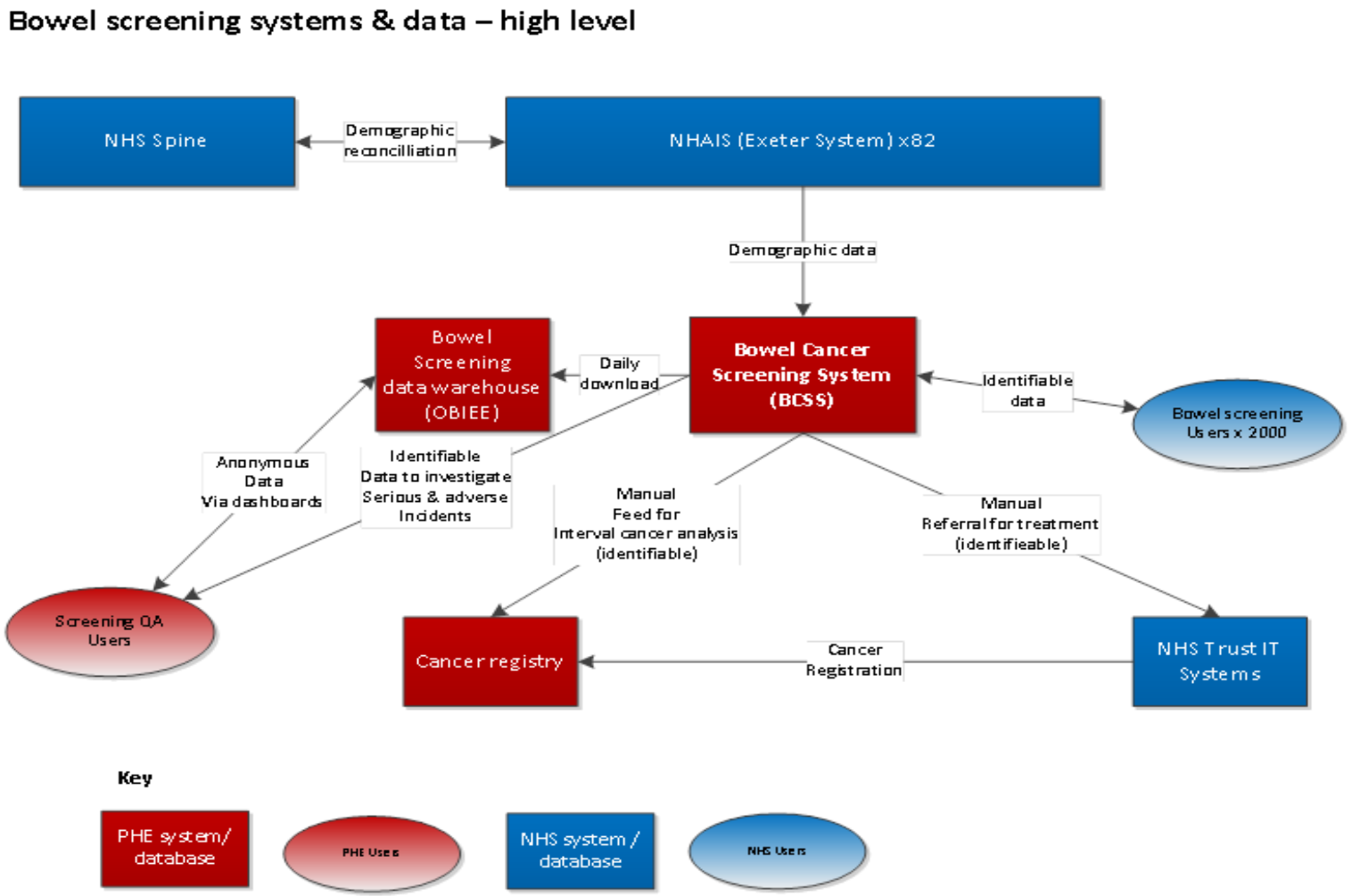


Figure 1: Data Schematic of the Bowel Cancer Screening System and the links with other data services from Public Health England. Data schematic was provided from Suzanne Wright, personal communication, Public Health England, with thanks.

1.4 Risk Scoring Systems for Colorectal Cancer using Electronic GP Records

One of the barriers identified to using a risk scoring system for screening is the collection of data from individuals; this can be facilitated by using routinely available data from electronic medical records. Risk prediction studies for colorectal cancer screening have suggested the possible use of electronic GP records to provide risk information.[10 11]

Electronic GP records have been used to develop the QCancer risk prediction model[12] which has been validated in women[13] and men[14]. This algorithm was derived from the QResearch database (incorporating electronic practice records) and takes into account both symptoms as predictors and other risk factors to determine an individual's absolute risk of colorectal cancer in a primary care setting. The algorithm was built using Cox's proportional hazards model on a derivation cohort to estimate risk factor coefficients (2,351,052 patients) and was tested on a validation cohort of 1,236,601 patients.[15] The predictors included in the final model were, age, family history, anaemia, rectal bleeding, abdominal pain, appetite loss and weight loss. The algorithm for men also included alcohol use and change in bowel habit. Both algorithms had good discrimination in the validation cohort (0.89 ROC for females and 0.906 for males) and the authors suggest this algorithm could be used to prioritise patients at sufficient risk for primary care referral. This model has been validated by Collins and Altman[12] with an AUC under the ROC curve of 0.91 for men and 0.92 for women.

Several studies have also looked at identifying and quantifying symptoms and diagnostic features of CRC.[16-21] NICE guidelines for referral, the CAPER scoring system and the Bristol-Birmingham (BB) equation can be used to identify patients who warrant further tests based on their symptoms and other variables.[20-22] The Bristol-Birmingham (BB) equation was derived using data provided by THIN and included the following variables: constipation, diarrhoea, change in bowel habit, abdominal pain, haemoglobin concentration, mean cell volume (MCV) and weight loss. The discrimination for the BB equation was 0.83 in the THIN dataset and 0.92 in the CAPER dataset. The CREDIBLE research project as an extension from this aimed to prospectively identify whether patients met NICE guidelines for referral by searching GP records.[23]

The models described above were developed for use in a primary care setting to identify those at sufficient risk for primary care referral. The current study however aims to develop risk prediction models using a BCSP cohort for referral decisions in a screening based setting.

1.5 Laboratory Parameters and Colorectal Cancer Diagnosis

Lab test results have been shown to relate to colorectal cancer prognosis and diagnosis.[24-27] A recent study using the THIN database and the Maccabi Healthcare Services (an Israeli dataset) combined blood measures, sex and age in a machine learning model (random forest model) to determine which individuals were at increased risk for colorectal cancer.[28] This model gave an AUC of 0.82. In addition, by combining the FOBT with the lab results and comparing it to the gFOBT alone, the model identified 48% more CRC cases.[28] This model was validated in a UK population using the CPRD database and gave AUC results comparable to the original Israeli model with an AUC of 0.776.[29]

Anaemia has been shown to predict colorectal cancer from blood count data present in electronic primary healthcare records.[25] Boursi *et al.*[30] developed risk prediction models for sporadic colorectal cancer using the THIN database. The AUC for a reference model of CRC risk factors was 0.58, for lab based only parameters 0.76 and for a model which combined both the reference model and lab parameters the AUC was 0.80.

The addition of lab based parameters based on these studies has shown improved predictive capabilities for colorectal cancer in a primary care setting. Studies identified in the systematic review reported in **Chapter 2,** which combine lab test results with the FIT, also show increased discriminatory power. A FOBT on its own, without other predictors may fail to detect intermittent bleeding or smaller lesions which may not bleed. The combination of abnormal blood results with the FOBT have been shown to improve sensitivity for detecting colorectal cancer.[31] The inclusion of these parameters within an algorithm for colorectal cancer screening referral warrants further investigation.

1.6 Survival Analysis

Time to event analysis or 'survival analysis' investigates the time to the occurrence of a defined event. This is usually time to death in clinical trials hence the term 'survival analysis'. At the end of the follow up period, not all individuals would have had the index event and so time to the event is unknown. This situation is referred to as right censoring of the data; the event (if it does occur) must occur to the right of the current or census time. For this reason, time to event data are usually skewed to early events and not normally distributed requiring specific methods for data analysis.[32] The semi-parametric Cox proportional hazards (PH) model is the most widely used for survival analysis and

estimates hazard ratios which can be used to measure how much a predictor increases/decreases the rate of the defined event.[33] The proportional hazards condition requires that explanatory variables are multiplicatively related to the hazard. The limitation of the Cox model is that it does not explicitly estimate the baseline hazard function.

The Cox model can be extended to provide an estimate for the baseline hazard but other parametric models (e.g. Weibull, generalised gamma, exponential models) also exist which provide estimates of the hazard and survival function for a combination of predictors. For parametric survival models the hazard is assumed to follow a specific statistical distribution.[34] Parametric proportional hazards models include the Exponential, Weibull and Gompertz models which produce hazard ratios like the semi-parametric Cox Regression model. Parametric survival models which follow the accelerated failure time metric include; Log-logistic, generalised gamma, Weibull and Log normal. The exponential of the coefficients produced by these models are time ratios instead of hazard ratios with a time ratio of >1 indicating that the predictor extends the time to the event and a time ratio of <1 indicating the predictor speeds up the time to the event.[34]

Although parametric survival models require the underlying hazard distribution to be specified, they are generally more informative than the equivalent Cox Regression model. This is because they provide the hazard function and therefore predicted survival probabilities for each patient can be derived. They are also more efficient and give more precise parameter estimates, provided that the underlying model is 'true'.[34]

Using time to event analysis for longitudinal health records is a more efficient use of data than looking at whether an event has occurred or not as seen in logistic regression analysis. In a longitudinal study, outcomes and exposures can occur at multiple time points for each participant.

Although studies usually investigate the time between response to treatment and recurrence or time from diagnosis to death, other studies have used survival analysis as a means for developing risk prediction models for cancer diagnosis.[13 14 35] Hippisley-Cox and Coupland[15] developed and validated a risk prediction algorithm for patients in primary care to facilitate early referral. The time to event used for this study was time to diagnosis of colorectal cancer or more specifically "the primary outcome was incident diagnosis of colorectal cancer recorded in the next 2 years"[15] (p. e29). This can make the model more future focused (or prognostic), rather than assessing accuracy at a fixed time point. The

paper also states that the Cox model was used to estimate the absolute risk of a patient in primary care currently having colorectal cancer.

1.7 Rationale

Risk prediction models combining predictors from electronic health records have shown promise in improving model and test accuracy for colorectal cancer detection. The richer data available from a GP record could be used to add a further dimension to a colorectal cancer screening model to improve discriminatory power and referral decisions. Lab test results available from GP records relate to the detection of colorectal cancer and may enhance the performance of risk prediction models as well as FOBTs which have greater sensitivity in men, can miss intermittent bleeding of lesions and have variable performance for cancers at different sites of the colon. To exploit longitudinal electronic GP data fully, different statistical methods in the form of survival analysis are required for censored data.

The aims of this study were therefore to determine (i) the availability of GP data for key predictors of colorectal cancer in the screening population using the THIN database and (ii) whether this additional information has the potential to be used to make more accurate screening referral decisions by developing multivariable risk prediction models.

Previous chapters have developed models based on a single time point to assess both model performance and test accuracy. This scenario translates well to using logistic regression with a binary endpoint at the time of screening. Survival analysis assesses the risk of disease over a certain time period (in this instance, diagnosis of colorectal cancer/polyps at 2 years). Given the longitudinal nature of the data, survival analysis is the most appropriate statistical method. Although this research uses a different modelling approach to previous chapters and uses the gFOBT as the screening test, the models can help to identify additional predictors which could be added to future risk based models for screening referral decisions. In addition, the methods used in this chapter can be applied for the FIT screening test when it is implemented in practice along with the specific clinical codes.

2.0 METHODS

Reporting will adhere to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[36] and the extended Reporting of studies Conducted using Observational Routinely collected Data (RECORD) guidelines[37].

2.1 Study population and data source

This research used the THIN database of anonymised GP records which has data for over 587 practices (5.67% coverage of UK in 2014) covering more than 12 million patients (including 3.6 million active patients).[3] THIN provides demographic information such as sex, age, Townsend deprivation score, diagnoses, symptoms and prescriptions recorded on GP records. The structure of the THIN database and its constituent data file types are described in detail in **Chapter 6**.

Risk prediction model development studies require large sample sizes in order to investigate the effect of a set of potential predictors on the outcome of interest and to produce a generalisable and valid model. Apart from the large source of data, further advantages of the THIN database are that it is UK wide, generalisable to the UK population [38] and allows patients to be followed up over time from an index event/date.

The BCSS currently sends FOBT results electronically to GP practice systems. The interconnectivity between GP records and screening records is likely to improve in the near future. Consequently, THIN will allow investigations into the predictors/factors which could be used to assist with referral decisions in practice.

THIN was used to identify men and women eligible for screening events for bowel cancer aged 60-74 years of age with at least a years worth of health records before being invited to take part in the BCSP. Practices are only included if they receive electronic notifications from the BCSP. Patients were followed from a year before they took the test (to ensure adequate symptomatic information to be identified) up to the point of diagnosis or a 2 year follow up to ensure any interval cancers/diagnoses are identified.

2.2 Study Design

The study design was an observational cohort covering the period 1st May 2009 to 17th January 2017. The **period of interest** was defined in part by the Acceptable Electronic BCSP (AEB) date for each practice as an assurance of data quality. This date is akin to the acceptable mortality reporting (AMR), before this date practices may not have routinely recorded patient deaths and de-registrations. [39] The purpose of this date was to ensure that the BCSP events recorded are those which are automatically sent by the screening programme to GP practices. Before electronic BCSP notifications were sent, practices had to manually input test results which may lead to biased recording since positive results may be more likely to be recorded. The development of the AEB date is reported in **Chapter 6**.

In brief, the incidence of BCSP FOBT Read codes for people aged 60-74 for each practice in England was examined from the additional health data (AHD) file in THIN (where lab results are recorded). This revealed a time-point at which the practice started to receive electronic notifications from the bowel cancer screening system. The AEB date was evaluated first, before identifying the observation period for each patient. Read code lists defining a BCSP notification for the English population are given in **Table 1**.

Eligible THIN practices included those practices with an AEB date and practices based in England (due to differences in bowel cancer screening systems and coding for Scotland, Northern Ireland and Wales). The eligible time period for each practice was the latest of the following; one year after the Vision practice software installation (to ensure the computer system was being fully utilised), AEB date (to ensure electronically received FOBT results are used for analysis) and AMR – to ensure the practice is routinely recording deaths).

Eligible patients were those aged between 60 and 74 with a BCSP notification during the period of interest. This generated a dataset of participants who had been adequately screened (with a positive or negative FOBT). For quality assurance, patients without a registration date, patflag A or C and regstat of 01, 02, 05, 99 (markers of data quality) were removed from the analysis (the THIN Researcher Guide provides more detail[4]). Patients without a Townsend score (based on four variables; unemployment, non-car ownership, non-home ownership and household overcrowding) or where the patient start date was greater or equal to the end date were also removed. Finally any patients with a previous

CRC diagnosis (before the most recent FOBT) along with any patients with a diagnosis of a high risk condition (hereditary nonpolyposis colorectal cancer (HNPCC) or familial adenomatous polyposis (FAP)) were excluded from the analysis. Practice and patient exclusion/inclusion criteria are shown in **Table 2** (greater detail is given in **Appendix 2**).

Entry into the cohort is defined as the latest of registration date plus one year (e.g. to allow symptoms to be recorded), AMR date, AEB date, Vision date plus one year, with a BCSP FOBT result. Exit from the cohort was the earliest of date died/left practice and the last practice collection date.

Practice Criteria	Patient Criteria
Practice Start Date: The latest of AMR, AEB date (acceptable electronic BCSP date) – defined by researcher) and Vision date plus one year.	Patient Observation Start: Later of Practice start date, registration date plus one year and age 60.
Practice End Date: The last collection date from each practice	Patient Observation End: Earlier of Practice end date, De-registration or death and age 74.
Exclude practices not in England	Period of interest Start: Latest FOBT recorded
Exclude practices where the practice start date is greater than or equal to the practice end date	Period of interest End: Earlier of patient observation end, or diagnosis of colorectal cancer/polyp within 2 years (most severe outcome and corresponding date used for analysis)

Table 2: Practice and patient inclusion and exclusion criteria.

2.3 Sample size

Feasibility data were obtained before applying for ethical approval in order to determine the level of BCSP activity recorded on the THIN database (**Table 3**). In terms of THIN data in 2013, around 36% of patients who are eligible for the BCSP had some evidence of documentation of their participation in the BCSP. At least 50,000 had documentation from 2010 onwards.

Year	Total Population of Practices (pyr)	1_Normal Result	2_Abnormal Result	3_Uncertain Result	4_Incomplete Participation	5_No Response	Total	Proportion with evidence of BCSP activity
2010	317529.14	5110	143	32	525	3609	9419	0.03
2011	314967.94	31379	608	1	1090	18078	51156	0.16
2012	306224.61	46229	733	4	1664	27272	75902	0.25
2013	279535.13	60760	1005	1	1075	37537	100378	0.36

Table 3: Count of Patients with Evidence of Screening Programme Activity. In 2013, 89% (307/346) of THIN practices in England had some evidence of BCSP activity recorded on their system; this was 74% (270/366) in 2012, 55% (209/380) in 2011 and 31% (124/393) in 2010.

Statistical power calculations for survival analysis relate to the number of events.[40] Some suggest that for reliable predictions, at least 10 events per variable/predictor need to be considered for analysis as demonstrated in simulation studies.[41-43]. More specifically, this is the number of cancers/polyps diagnosed per degree of freedom. Categorical variables have (n-1) degrees of freedom and continuous variables have one degree of freedom.

The dataset derived for the multivariable modelling analysis had 1,197 colorectal cancer and polyp diagnoses (sample population = 98,303) and considered 33 degrees of freedom (45 including interactions) giving 36.27 events per variable or 26.6 including interactions. For the model with negative FOBTs only, there were 587 events (sample population = 95,792) and considered 31 degrees of freedom (39 including interactions) giving 18.94 events per variable or 15.05 including interactions.

2.4 Ethical Approval

Ethical approval was given by the Scientific Review Committee (SRC) administered by IMS Health (SRC Reference Number: 16THIN037 Date: 26/05/2016). See **Appendix 3**. This is an independent scientific review committee which can grant ethical approval for a study using data from THIN which is pseudonymised and pre-collected. The NHS South-East Multi-centre Research Ethics Committee approved THIN data collection in 2003 for this purpose.[44]

2.5 Model Outcome and Index Date

The index date used for survival analysis was the latest BCSP FOBT result electronically sent to GP practices. The index date for the FOBT ranged from 13th May 2009 to 5th January 2017. The outcome was the diagnosis of colorectal cancer/polyps after this index date for a maximum follow up of 2 years. Two years was chosen for follow up as this represents one screening round in the NHS and existing cancers are likely to be clinically identified within this period.[15]

Colorectal cancer and polyps are defined using the Read code lists developed and supplied in **Chapter 6**. If both polyps and colorectal cancer had been diagnosed within the 2-year period of follow up then the most severe outcome was used for analysis. Polyps were

included in the outcome as during screening the aim is to detect early stage lesions before they develop into later stage cancers as well as more severe cancers.

2.6 Model Predictors

Predictors included in the analysis were identified from previous studies incorporating symptoms and lab test results as well from the systematic review reported in **Chapter 2** and NICE guidelines (see **Table 4**).[18 21 45-47] These predictors were measured at the time of entry to the study/up to 365 days before the index date to ensure they were associated with the outcome. The full specification of these variables are provided in **Appendix 4** and the table contents for data extraction and analysis in **Appendix 5**. For multivariable analysis, 29 variables were considered during the modelling process. Rectal bleeding, abdominal mass and abnormal rectal examination are considered red flag symptoms and the patients would have been referred in the primary care setting so were not included in the multivariable analysis.

Methods to extract AHD variables (often based on Read code lists as well as AHD codes) are given in **Chapter 6**. Drug codes for three types of drug (anti-motility drugs, antispasmodics and laxatives) were derived in part using the methods of Dave and Petersen[48] and the THIN Data Guide for Researchers[4]. All the variables were recorded before the index date (date of FOBT) since at the time of deciding referral, this would be the only information available. Anything recorded after this date would not be able to contribute to the risk prediction model.

Study/Guidelines	Included Variables
NICE Guidelines (NG12) Suspected cancer: recognition and referral[45]	1.3.1 Two week referral for suspected colorectal cancer:[45] 1. Adults aged 40≤ with unexplained weight loss and abdominal pain 2. Adults aged 50≤ with unexplained rectal bleeding 3. Adults aged 60≤ with: iron-deficiency anaemia or changes in their bowel habit 4. Tests show occult blood in their faeces Consider a suspected cancer pathway referral for:[45] 1.3.2 Adults with a rectal or abdominal mass 1.3.3 Adults aged under 50 with rectal bleeding and any of the following: abdominal pain change in bowel habit, weight loss, iron-deficiency anaemia. 1.3.4 This recommendation has been superseded by the one below using FIT **DG30.**[49]
NICE Diagnostic Guidance (DG30) Quantitative faecal immunochemical tests to guide referral for colorectal cancer in primary care[49]	In addition, new guidance using quantitative FIT to guide referrals suggest use of FIT in people without rectal bleeding who have unexplained symptoms but do not meet the criteria for a suspected cancer pathway. A threshold of 10 micrograms of haemoglobin per gram of faeces to be used for referral.[49] Variables: Abdominal pain, weight loss, change in bowel habit, rectal bleeding, iron-deficiency anemia, anemia in the absence of iron deficiency, rectal/abdominal mass, FOBT, FIT
Hippisley-Cox and Coupland[15]	Algorithm for females included: age; family history of gastrointestinal cancer; anaemia; rectal bleeding; abdominal pain; appetite loss; and weight loss. Algorithm for males same as above but includes alcohol use and change in bowel habit.
Marshall *et al.*[21]	Final model included: Constipation, diarrhoea, change in bowel habit, abdominal pain, haemoglobin concentration mean cell volume and weight loss. In univariable analysis, a positive FOBT, abnormal rectal examination and abdominal mass had a strong association with a diagnosis of colorectal cancer.
Hamilton[20] (CAPER Studies)	Variables independently associated with colorectal cancer after multivariable analysis included: Rectal bleeding, loss of weight, abdominal pain, diarrhoea, constipation
Hamilton *et al.*[46]	Final multivariable analysis – rectal bleeding, change in bowel habit, abdominal pain, diarrhea, constipation, weight loss, haemoglobin concentration, mean red cell volume.
Lab Measurements Hamilton[20]	Abnormal primary care investigations: Positive FOBT, haemoglobin concentration, blood sugar levels
Lab Measurements Boursi *et al.*[30]	Lab based model only: Hematocrit, mcv, neutrophil-lymphocyte ratio, lymphocytes, creatinine, BUN Multivariable model based on all variables included: Sex, haemoglobin, MCV, white blood cells, platelets, NLR, prescription of metformin/oral hypoglycemic medications
Lab Measurements Marshall *et al.*[21]	MCV, haemoglobin concentration
Lab Measurements, Kinar *et al.*[28] validated by Birks *et al.*[29]	Age, sex, blood count data including; Haemoglobin, mean corpuscular haemoglobin, haematocrit, platelets, mean cell volume, mean corpuscular haemoglobin concentration, red blood cell distribution width, lymphocytes(%), red blood cell count,, mean platelet volume, neutrophils(%), neutrophils (number), monocytes (number). Also investigated combining FOBT result.

Table 4: Predictors identified from previous risk prediction model studies and NICE guidelines

Symptoms used for analysis were the latest record prior to the index date (date of latest FOBT) within 365 days to ensure that a record is more likely to be associated with the outcome (i.e. diagnosis of colorectal cancer/polyp). These symptoms included, abdominal pain/antispasmodic drug use, abnormal rectal examination, constipation/laxative, diarrhoea/anti-motility drug use, change in bowel habit, flatulence, loss of appetite, rectal bleeding/melaena, tiredness, weight loss. These were coded as binary variables.

As identified in previous studies,[21] Drug codes for anti-motility, laxatives and antispasmodic prescriptions were used as a proxy for the following symptoms respectively: diarrhoea, constipation and abdominal pain. If the hazard ratios were similar in univariable analysis,

the corresponding prescription and symptom were combined for multivariable analysis which was the case for antispasmodic drugs and abdominal pain.

Diagnoses or other conditions used for analysis included, venous thromboembolism (which includes pulmonary embolism and deep vein thrombosis) up to 365 days prior to the index date, ulcerative colitis, IBS, Crohn's disease if ever recorded (prior to the index date), diabetes up to 365 days prior to the index date, polyps and diverticulitis if ever recorded before the index date.

Lab tests/other investigations included; platelet count, haemoglobin concentration, ferritin and MCV the latest record up to 365 days prior to the index date. All lab results recorded for 2 years prior to the index date were also extracted so they could be investigated longitudinally. The percentage difference between the last two recorded values before the index date was used to define new variables to analyse individual longitudinal trends. An indication of anaemia is incorporated in the haemoglobin concentration variable. Thrombocytosis is also captured using platelet count in the AHD records. Low ferritin was investigated in the same way, using ferritin levels recorded in the AHD records. The latest height, weight, BMI records before the index date were used as well as all the records for individuals two years prior to the index date.

A FOBT performed in primary care up to 365 days prior to the index date was included as a potential variable. The BCSP latest FOBT result was extracted for the index date. All previous BCSP FOBT results were also extracted in order to have an individual's screening history. Depending on the result, number of previous negative FOBTs and number of previous positive FOBTs were defined as additional variables.

Other factors included, family history of gastrointestinal cancer if ever recorded before the index date. The last recorded entry prior to the index date for alcohol consumption, smoking status, blood group, Townsend score (a measure of social deprivation) and ethnic group.

A BCSP flexible sigmoidoscopy recording any time before the index date; this is usually a one off test at age 55 which has been recently introduced and rolled out in England as part of the BCSP. When investigating this further, there were not enough recordings for this test available from the dataset so it was dropped from multivariable analysis.

Ethnic group was not included as a candidate variable as there is evidence to suggest currently that the recordings for some groups are not representative of the UK despite the

introduction of Quality and Outcomes Framework (QOF) incentives.[50] Other factors such as alcohol consumption and smoking status have been shown to be well recorded in THIN.[51] In addition, abnormal rectal examination was removed from model development since there were very few records which meant the model did not always converge. This is a red flag symptom and the individual would have been referred under the two week referral scheme. There were more records after the index date but the recording of this symptom would have suggested a procedure had been performed.

2.7 Statistical Analysis

The following analyses were undertaken:

i) Estimate of the sensitivity and specificity of the FOBT for colorectal cancer and polyps using the derived screening population.
This helped to determine if the data extracted based on the electronic BCSP notifications was valid if it gave similar results to that reported in the literature.

ii) Univariable analysis (Cox Regression) and analysis of the availability of variables in the screening population.
This analysis assessed the completeness of variables which may be useful in a risk based prediction model and identified predictors with a strong independent association.

iii) Kaplan Meier Survival Analysis:
A) Survival analysis based on time to diagnosis and time to death from a positive or negative FOBT (using the whole derived screening cohort); Kaplan-Meier estimates stratified by FOBT result and sex.
B) Survival analysis based on time to diagnosis and time to death for patients with a negative FOBT only; Kaplan Meier estimates stratified by sex.
C) Survival analysis time to diagnosis and time to death stratified by true negatives, true positives, false negatives and false positives for a population with either a diagnosis or a minimum of 2 years of follow up information.
These analyses described and checked the validity of the data extracted for analysis and identified predictors which affect survival by comparing the survival functions of different groups/covariate patterns.

iv) Development of a risk prediction model for those with a positive or negative FOBT using Cox Regression.

This analysis identified additional predictors from a multivariable model which have the potential to be added to a risk based model to make more accurate screening referral decisions.

v) Absolute risk probabilities (survival predictions) were derived using the Cox regression model by estimating the baseline survival at 2 years.
This analysis was performed to determine individual risk probabilities from the previous model and determine the distribution of risk in the sample population based on the predictors in the multivariable model.

vi) Assessment of parametric survival models using AIC, cumulative hazard plots, Kaplan Meier function plots and Cox-Snell residuals for those with a positive or negative FOBT.
Parametric models were investigated as an extension to Cox Regression to determine whether these models gave a better fit to the data and therefore more accurate parameter estimates.

vii) Development of a risk prediction model for those with just a negative FOBT using Cox Regression to assess which other predictors could be predictive of cancer and warrant referral (false negatives, true negatives).
This analysis allowed the identification of additional predictors which could be used in a screening population with negative results for screening referral decisions.

viii) Absolute risk probabilities (survival predictions) were derived using the Cox regression model by estimating the baseline survival at 2 years.
This analysis was performed to determine individual risk probabilities from the previous model and determine the distribution of risk in the sample population based on the predictors in the multivariable model.

ix) Nomogram of the risk prediction model.
A Nomogram for the model developed in (vii) was produced as an alternative method of presenting the risk equation.

x) Assessment of parametric survival models using AIC, cumulative hazard, survival and Cox-Snell residuals for those with a negative FOBT.
Parametric models were investigated as an extension to Cox Regression to determine whether these models gave a better fit to the data and therefore more accurate parameter estimates.

'Survival' when assessing time to diagnosis in these analyses refers to colorectal cancer/polyp free survival. 'Survival' when assessing time to death refers to overall survival.

2.7.1 Test Accuracy

Using the derived cohort, this analysis was restricted to patients who had either the outcome within 2 years or at least 2 years of follow up information after the FOBT (index test) n=32,004. Two by two tables were constructed to estimate sensitivity, specificity, PPV (positive predictive value) and NPV (negative predictive value) of the FOBT. The diagnostic outcome was broken down by colorectal cancer diagnosis or polyp diagnosis – with the most severe outcome used for analysis if both were recorded during the time period.

2.7.2 Univariable Analysis and Data Missingness

Cox regression was used to estimate hazard ratios for all considered predictors. Hazard ratios measure how much a predictor increases or decreases the rate of a defined event.[33]

The level of complete/missing data was also recorded in order to determine which predictors could be included in the multivariable analysis. Ideally, variables which are readily available from GP records for the screening population and variables which are strong predictors of colorectal cancer/polyps are the most useful for CRC screening decisions.

2.7.3 Kaplan-Meier Estimations – Time to Diagnosis (colorectal cancer/polyp free survival) and Time to Death (survival)

Two useful measures for describing survival data are (i) the survivor probability or survivor function *S(t)* and (ii) the hazard h(t).[32] In this setting, the survivor probability refers to the probability that an individual is free from colorectal cancer/polyp diagnosis for time to diagnosis. For time to death the survivor probability refers to the probability that an individual survives/is alive. The hazard is the probability that the patient has an event at a particular time point. The nonparametric Kaplan-Meier survivor estimate was used for *S(t)* and the Nelson-Aalen estimator for *H(t),* the cumulative hazard.[32]

Kaplan-Meier survivor estimates were used to estimate the probability of survival at particular time points for a sample population with positive and negative results and for a

sample with just negative results.[52] Time to diagnosis was the time from a FOBT result to CRC/polyp diagnosis within 2 years. To compare survival between groups of patients, the non-parametric log-rank test was used to compare sex and FOBT result. The log-rank test compares the groups using the ordering of failure times. Time to death was the time from a FOBT result to any cause of death.

In addition, the colorectal cancer free survival and survival between patients with a true positive (TP), true negative (TN), false positive (FP) and false negative (FN) result were plotted on a Kaplan-Meier curve and compared using a log-rank test.

This analysis allowed the data extracted for analysis to be described and checked for validity and also identified subgroups which affect colorectal cancer/polyp free survival and overall survival.

2.7.4 Model Development Strategy

Two different models were produced using Cox Regression. The first was for a population with both positive and negative FOBT results. The second model was for a population with negative FOBT results only to determine whether other factors could be used to decide whether a patient should be referred if they have had a negative result.

The model development strategy was dependent in part on the level of missing data for certain predictors. Variables which had >60% missing data were excluded from the analysis since the sample size would have been restricted too much keeping the other factors in the analyses. This approach was a compromise between including important lab parameters and retaining the sample population for model development. Initial analysis investigated the level of missingness for each variable, and the varying ways in which it could be recorded (AHD versus Read code versus combined). For example, alcohol consumption could be used as a Read code in the medical table or taken from the AHD tables as units per week. Where appropriate, the continuous versions of a variable if well recorded, were used for model development. For instance, thrombocytosis can be captured using the platelet count measured in the AHD table. Colorectal cancer has been shown to be more frequently diagnosed in those with thrombocytosis. [27]

The TRIPOD guidelines recommend using a continuous variable rather than dichotomising into different groups as this loses additional predictive information.[53] However, since setting a cut-off for different blood parameters is also clinically meaningful as it can

indicate underlying disease, categorised blood measurements were considered for, platelet count, ferritin, haemoglobin concentration and mean cell volume.

Symptoms: Weight loss, abdominal pain combined with antispasmodic prescription, constipation, diarrhoea, change in bowel habit, loss of appetite, tiredness, flatulence.

Lab test results: Haemoglobin category, MCV category, platelet count category.

Screening Utilisation: Latest FOBT result, previous negative FOBT result from BCSP, previous positive FOBT result from BCSP, previous polyp recording before index date, Primary Care FOBT.

Co-morbidities and previous diagnoses: Crohn's disease diagnosis, ulcerative colitis, diabetes, IBS, diverticulitis, venous thromboembolism.

Demographic Factors: Sex, age at FOBT.

Lifestyle and anthropometric factors: Smoking status, BMI, alcohol consumption (number of units consumed a week).

Additional Factors: Townsend quintile, family history of gastrointestinal cancer.

Cox regression using Efron's method of handling ties[54] was used to develop a multivariable model. Multivariable fractional polynomials (MFPs) were applied with the methods implemented in Stata[55] to model non-linear relationships with continuous predictors.[55] This allows the variable to be kept continuous and allow for some form of non-linearity.[56 57] It is best practice to assess non-linear functions after adjusting for other predictors in the model. Fractional polynomials provide more flexible parameterisation for continuous variables compared with regular polynomials by providing a richer class of possible functional forms to fit to the data.[56] MFPs also give a better fit to real life data, capture more complex associations with the outcome and can lead to better risk predictions.[56]

The 'mfp' function in Stata selects the variables which best predict the outcome, more details about the selection algorithm are supplied in the associated Stata 'mfp' manual and journal articles.[55 56] In brief, this method combines backwards elimination with a search for the most suitable FP transformation of a continuous predictor. For backwards elimination a p-value of 0.05 was used to determine whether to keep a predictor in the model (a variable is removed if dropping it from the model causes a nonsignificant increase in the deviance).[55] P-values for testing between fractional polynomial models was set at 0.05.

Interactions were also assessed using a p value of 0.05. Interactions investigated included age and sex, FOBT result and sex, FOBT result and age, MCV and age, MCV and sex, alcohol and sex, alcohol and age, smoking and age, smoking and sex as these interactions/associations have been found to be significant from the literature. A likelihood ratio test was carried out to determine whether the model with interactions provided significantly better fit than a model without interactions. When reporting the final model the Cox Regression coefficients are provided along with bootstrapped standard errors (100 replications).

Since parametric models can be more efficient than the equivalent Cox regression model by producing more precise estimates and offering more with post-estimation such as predicted survival probabilities, hazard and predicted times, parametric survival models were also investigated. Parametric proportional hazards models are the exponential, Weibull and Gompertz models. These models have a monotonic hazard function which increases or decreases over time whereas the loglogistic, generalised gamma and lognormal models have hazard functions with turning points.[33] The fit of the parametric models were compared using the AIC, cumulative hazard plots (where the cumulative hazard is plotted and parametric curves fitted to it), Kaplan Meier Plots (where the Kaplan-Meier survivor function is compared with the function predicted if the survival times followed a particular parametric distribution) and Cox-Snell residuals to determine the model with the best fit.

A summary of the models investigated and the sample population used to derive the models is given below in **Table 5.**

Model Population	Model
Positive and Negative FOBT Results (n= 98,303)	Cox Regression
	Parametric Models
Negative FOBT results only (n=95,792)	Cox Regression
	Parametric Models

Table 5: Model types investigated for multivariable analysis

2.7.5 Model Performance

The model performance was assessed using Harrell's C statistic which takes into account the censored nature of survival data.[58 59] This is used as a measure of discrimination akin to the AUC ROC for the Cox Regression models. Harrell's C statistic measures the agreement of predictions with observed failure order and is the proportion of subject pairs where the predictions and outcomes are in agreement. [60] The values range between 0 and 1. Somers'

D rank correlation which is related to Harrell's C statistic (D=2(C-0.5)) ranges from -1 to 1 was also reported. [60 61] The D statistic, R^2 and adjusted R^2 were also determined as further measures of discrimination performance. Separation was also visually assessed by plotting Kaplan-Meier survival curves for 4 risk groups (where the linear predictor is divided into 4 groups based on Cox's method).[62] Five or fewer groups have been recommended for this purpose.[63 64] This also reflects the underlying C statistic, D statistic and R^2 values.

The linear predictor and its distribution was summarised by plotting a histogram of values and reporting the corresponding mean and standard deviation. The linear predictor is the linear combination of the predictors in the model with their beta coefficients but without the intercept or baseline hazard (LP = $\beta_1 x_{1+} \beta_2 x_{2+} \beta_3 x_{3\ldots}$). Analysing the distribution of risk in the population is useful when applying the model to another dataset for external validation to explain any underlying differences in performance and to assess the usefulness of the model.[65 66] The spectrum effect relates to the underlying case mix from which the model is derived. The performance of tests/models varies among different population subgroups and this must be considered when externally validating a model.[67]

Optimism of the model was assessed by calculating the heuristic shrinkage factor of Van Houwelingen (given by: (chi^2 – degrees of freedom)/chi^2).[68] This shrinkage factor was then applied to the linear predictor to generate a shrunken linear predictor adjusted for optimism and the corresponding distribution (mean and standard deviation) was also reported.

To adjust performance statistics for optimism, internal validation was performed using 100 bootstrap replications for the C statistic, c-slope, D statistic and R^2 to quantify optimism. The optimism was then subtracted from the apparent performance statistics to give optimism adjusted values for these performance parameters. The bootstrap uniform shrinkage factor was then reported (based on the optimism adjusted c-slope value) as an alternative to the heuristic shrinkage estimate.

Calibration of the models was assessed by plotting a calibration curve of observed probability versus expected probability for deciles of risk. This graphical approach also allows discrimination to be assessed visually by investigating the spacing between risk groups. The calibration slope is reported once adjusted for optimism.

Differences in performance between parametric models was assessed using the Akaike Information Criterion (AIC) [69] and Bayes Information criterion (BIC). [70] The AIC is defined as

the deviance (2 times the minus maximised log likelihood) plus 2*k* where *k* is the number of fitted parameters.[33] The BIC is the deviance plus *k*logn where n is the number of events and *k* is the number of model parameters. Since the AIC and BIC for the Cox model are calculated using partial likelihood, the Cox and parametric models could not be directly compared using these statistics. [33] The goodness of fit was assessed for the Cox Regression and parametric models using Cox-Snell residuals. If the model fits the data well then the cumulative hazard of the Cox-Snell residuals should be a straight line at 45 degrees. [60] The goodness of fit is assessed by investigating whether the model represents the survival patterns of the data adequately.[40] Cumulative hazard and survival plots were also used to compare the fit of the models.

2.7.6 Absolute Risk Probabilities

Predicted probabilities of colorectal cancer/polyps were derived for each patient and their covariate pattern. The shrunken linear predictor was used to estimate a new baseline survival (adjusted for optimism) which was estimated non-parametrically at 2 years. The shrunken linear predictor was combined with the baseline survival to generate individualised risk predictions. In order to obtain an event probability as opposed to a survival probability, the result of this was subtracted from 1 to generate the probability of colorectal cancer/polyps being detected over a 2 year period.

Cox regression is a semi-parametric model and therefore does not give the baseline hazard $h_0(t)$ or survival $S_0(t)$ and so this was estimated using non parametric methods at a particular time point. Non parametric estimation of the baseline survival was obtained using a zero covariate value and the methods implemented in Stata. Baseline survival from a particular time-point (2 years) can then be obtained from a Kaplan-Meier curve or accompanying results. See **Equation 1** for the Cox Regression equation to estimate survival probabilities.

$$S(t) = S_0(t)^{\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots)}$$

S_0 = Baseline survival probability at 2 years

Equation 1: Risk equation for determining corresponding probabilities from the Cox Regression model.

Individualised risk predictions can be used for decision making by setting a probability cut point of an individual being diagnosed with cancer in the next 2 years (representing a round of screening). The risk prediction model can subsequently be validated and applied in a screening population as a screening test to determine if there is an improvement in cancer detection or sensitivity/specificity.

The model for the population with negative tests was displayed as a Kattan style Nomogram using a Stata Nomogram generator.[71] This allows the probability to be determined graphically as an alternative to using the risk equation to generate probabilities.

2.7.7 Cox Regression Diagnostics

To test the proportional hazards assumption of the model, Schoenfeld residuals of the covariates were examined. For covariates with a p value of less than 0.05 (testing for the null hypothesis of a nonzero slope) the scaled Schoenfeld residuals were plotted. A straight horizontal line supports that there is not a violation of the proportional hazards assumption. Log-log plots were also plotted for these variables to assess proportionality by determining whether the lines are roughly parallel. The overall fit was assessed using Cox-Snell residuals by plotting the Nelson-Aalen cumulative hazard function against Cox-Snell residuals and assessing the corresponding fit. If the model fits the data well then the cumulative hazard of the Cox-Snell residuals should be a straight 45 degree line which reflects an exponential distribution with hazard equal to 1 across time (t).[60]

2.7.8 Parametric Survival Models

Parametric models were investigated as an extension to Cox Regression to determine whether these types of model gave a better fit to the data and therefore more accurate parameter estimates. The generalised gamma, loglogistic and log normal parametric models use the accelerated time metric and allow derivation of time ratios which can be seen as more interpretable than hazard ratios.[34] The Weibull, exponential and Gompertz parametric models on the other hand have a proportional-hazards parameterisation. The AIC was used to compare the different parametric survival models since the models do not require to be nested as with likelihood ratio testing. The fit was investigated further by plotting Cox-Snell residuals, Nelson Aalen cumulative hazard plots and Kaplan Meier function graphs for the parametric models to assess the fit visually.

Calibration and discrimination were reported for the best fitting parametric models according to the above investigations (residual plots, Kaplan-Meier function and Nelson Aalen cumulative hazard plots as well as the AIC). Harrell's C-statistic along with the Calibration plots were reported for the generalised gamma model (the best fitting model for the scenario which included FOBT results) and the Gompertz model (the best fitting model for the scenario including participants with negative FOBTs only). The 'somersd' package available in Stata can determine Harrell's C and Somers' D statistics and provide confidence intervals.[72] This package can be extended for use in parametric models and so was also used to determine Harrell's C statistic for models fitted with 'streg'. The R^2 used in this instance was Royston and Sauerbrei's (2004) R^2_D measure of explained variation for survival models based on their index of discrimination (D).[73] The adjusted R^2 measure also considers the number of covariates in the model. For non-proportional hazards models R^2 for explained variation is not interpretable but can be used as an index of determination.[74]

As an extension to the generalised gamma model, a Wald test was performed testing the hypothesis that kappa is equal to zero which would have indicated a lognormal model was an appropriate model for the dataset.[60 75] A Wald test for whether kappa was equal to one was also investigated which would suggest a Weibull model would be an adequate model. Calibration and discrimination were therefore also reported for these models in the scenario which included FOBT results in the model.[60 75]

3.0 RESULTS

3.1 Study Population

3.1.1 Overall Screening Cohort Derived from THIN

The study flow diagram is shown in **Figure 2**. The screening cohort defined from the THIN database gave 292,168 patients across 360 practices aged 60-74 with a positive or negative FOBT result. The cohort was 53.26% female, with a mean age of 66.43. Practices were restricted to England and those which receive electronic BCSP notifications. Those with high-risk conditions (e.g. familial adenomatous polyposis) were also excluded from the analysis.

The primary outcome investigated was both colorectal cancer and polyps combined. The most severe diagnosis within 2 years was colorectal cancer for 929 patients and polyps for 1960 patients (2889 total) (See **Table 6**). The number of patients who died during the 2 year follow up from the index test was 3169. The number who died without a time restriction (I.e. until the end of data collection) was 3,842.

	Colorectal Cancer	Polyp	Diagnosis not recorded
Most severe diagnosis within 2 years of index test	929	1960	289,279
Most severe diagnosis to the end of follow up from the index test	1024	2158	288,986
Number of CRC within 2 years of index test	929	-	291,239
Number of Polyps within 2 years of index test	2131	-	290,037
Number of CRC to the end of follow up from the index test	1024	-	291,144
Number of Polyps to the end of follow up from the index test	2341	-	289,827

Table 6: Diagnostic outcomes within a 2 year follow up period and until the end of follow up for the cohort.

3.1.2 Test Accuracy Population

For test accuracy purposes, only those participants with a diagnosis within two years or with two year follow up were investigated (n = 32,004). For this population, there were 2,610 positive FOBTs and 29,394 negative FOBTs (positivity 8.16%). Participants were 51.46% female with a mean age of 66.08 years. There were 2889 colorectal cancers and advanced adenomas and 765 deaths.

3.1.3 Participants with Positive and Negative FOBTs

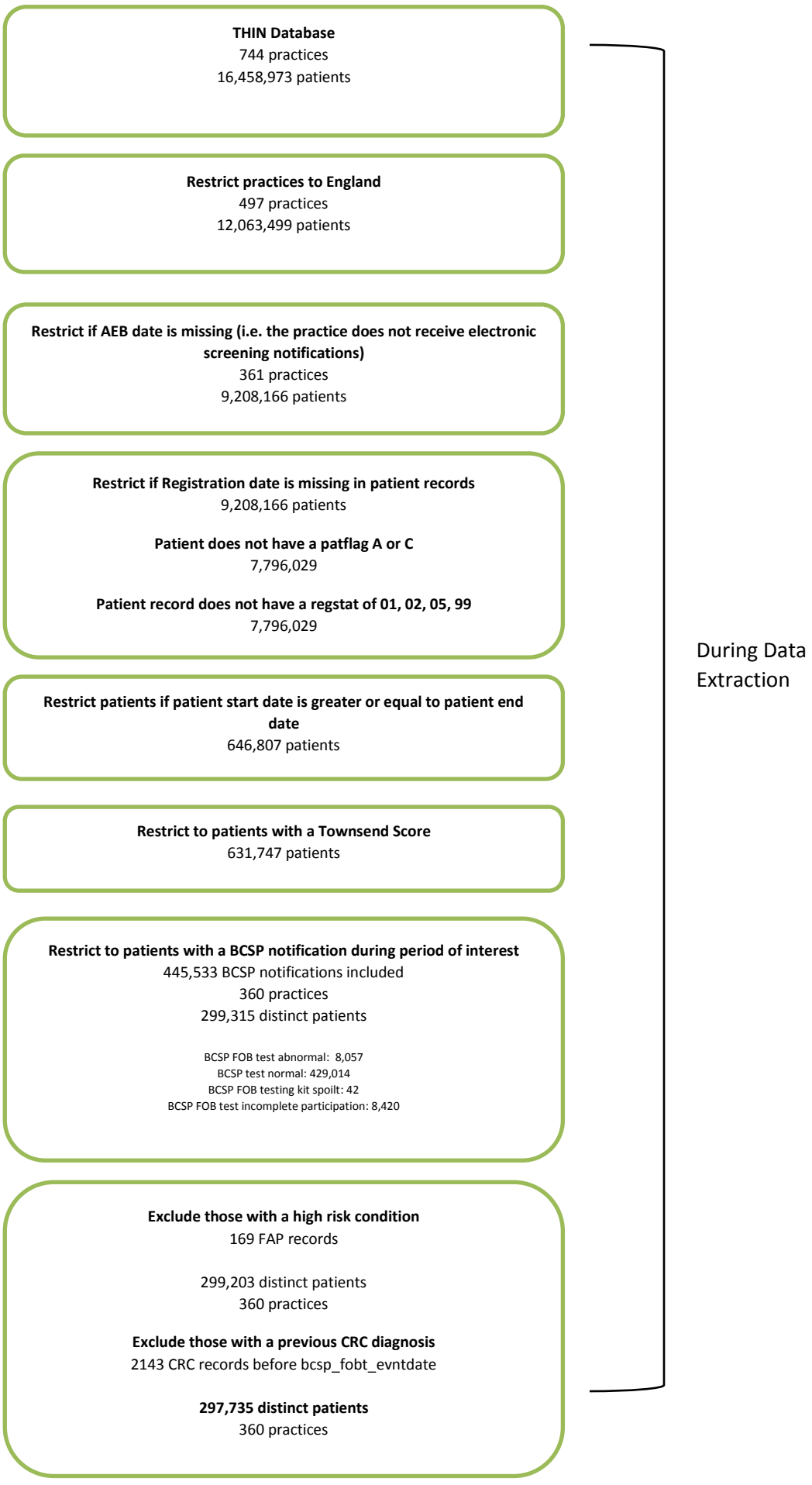
For multivariable analysis, complete cases of participants with the variables of interest were included (n=98,336). For this population, the mean age was 66.97 years and 50.04% were female. There were 2,511 abnormal FOBT results and 95,825 normal FOBT results. The number of cancers and polyps detected in this population were 350 and 847 respectively. After setting the data up for survival analysis, 33 observations occurred on entry and so 98303 observations remained with 1197 events and 38,005,604 person years.

The factors which limited the sample size for this investigation were the laboratory results; Hb concentration, platelet count and MCV. The cancer/polyp detection rate for those with a laboratory record (for all three results) was around 1.19% and those without 0.83% (Pearson's chi-squared p=<0.001) (**Table 7**).

	Haemoglobin Concentration		**MCV**		**Platelet Count**	
	Without Record	*With Record*	*Without Record*	*With Record*	*Without Record*	*With Record*
Cancer/Polyp	1345	1544	1,352	1,537	1,349	1,540
No Cancer/Polyp	160,784	128,495	161,293	127,986	161,091	128,188
Cancer Detection Rate (%)	0.830	1.187	0.831	1.187	0.830	1.187

Table 7: Cancer detection rates for participants with and without laboratory results (haemoglobin concentration, MCV and platelet count).

Participants with Negative FOBT Population (complete cases n=95,825): Patients with just negative FOBT results recorded as their latest screening test result were investigated in order to determine whether additional predictors could be used for referral decisions. The sensitivity of the FOBT could be improved by adding these additional literature-driven factors. For this population there were 144 cancers and 443 polyps diagnosed, 50.32% were male and the mean age was 66.97. After setting the data up for survival analysis, 33 observations occurred on entry and so 95,792 observations remained with 587 events and 37,154,249.5 total analysis time at risk and under observation.


THIN Database
744 practices
16,458,973 patients
Restrict practices to England
497 practices
12,063,499 patients
Restrict if AEB date is missing (i.e. the practice does not receive electronic screening notifications)
361 practices
9,208,166 patients
Restrict if Registration date is missing in patient records
9,208,166 patients
Patient does not have a patflag A or C
7,796,029
Patient record does not have a regstat of 01, 02, 05, 99
7,796,029
Restrict patients if patient start date is greater or equal to patient end date
646,807 patients
Restrict to patients with a Townsend Score
631,747 patients
Restrict to patients with a BCSP notification during period of interest
445,533 BCSP notifications included
360 practices
299,315 distinct patients
BCSP FOB test abnormal: 8,057
BCSP test normal: 429,014
BCSP FOB testing kit spoilt: 42
BCSP FOB test incomplete participation: 8,420
Exclude those with a high risk condition
169 FAP records
299,203 distinct patients
360 practices
Exclude those with a previous CRC diagnosis
2143 CRC records before bcsp_fobt_evntdate
297,735 distinct patients
360 practices
During Data Extraction

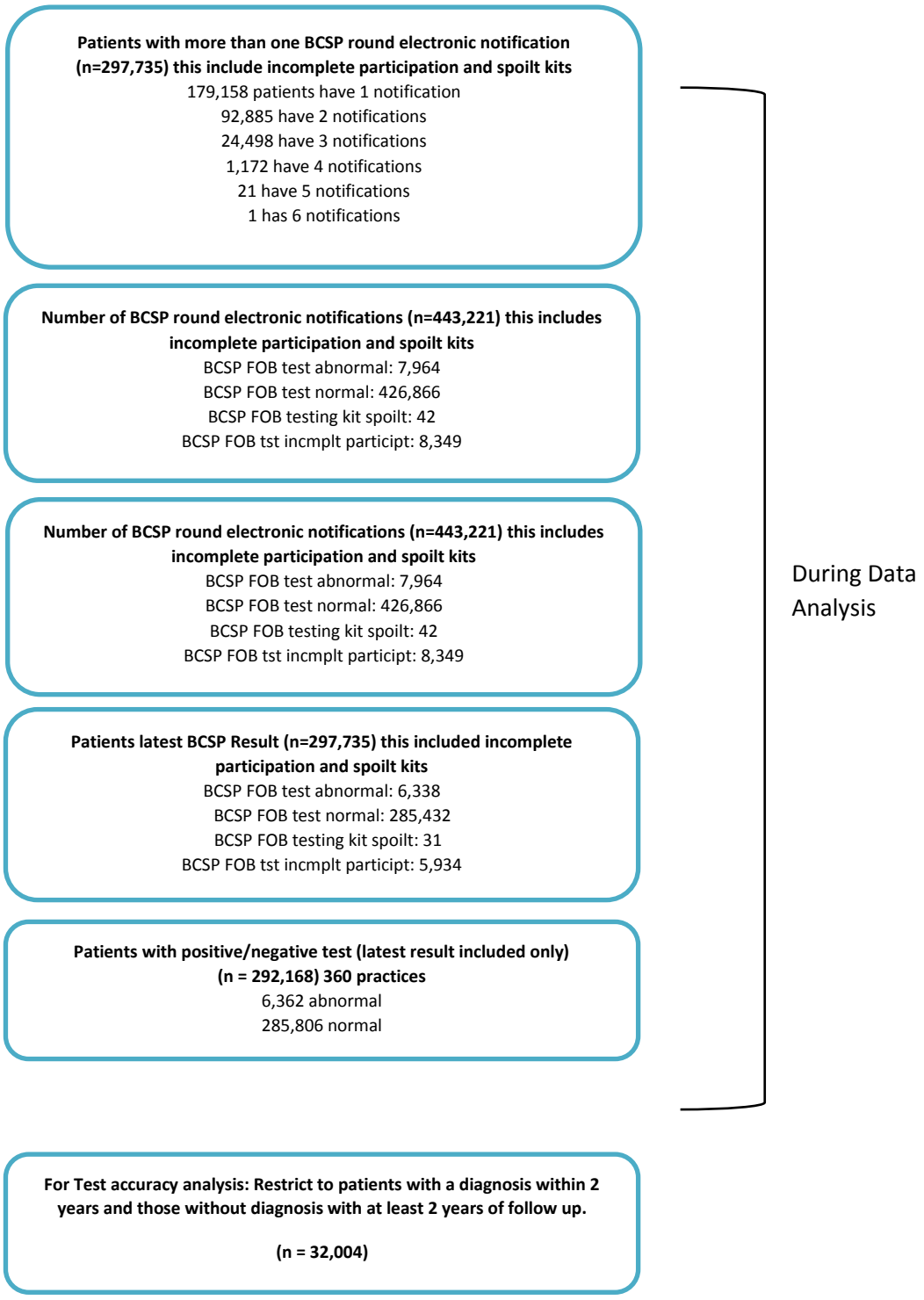


Figure 2: Study flow diagram for data extraction from THIN (1st plot) and for data analysis (2nd plot).

3.2 Test Accuracy

This analysis helped to determine if the data extracted based on the electronic BCSP notifications was valid by comparing to the FOBT results reported in the literature. There were 6,362 abnormal FOBT results and 285,806 normal FOBT results equating to a positivity of 2.18% out of those who had adequately participated. This is a similar positivity to that previously reported in the literature.[76] The analysis is restricted to patients who have either had the outcome within 2 years or have a minimum of 2 years follow up after the index test (n=32,004). The two by two table for this analysis is presented below (**Table 8**) providing a two-year sensitivity (screening round sensitivity) of 52.8% and specificity of 96.3%. The mean age at FOBT was 66.08 (SD: 4.34) with 15,535 males and 16,469 females (51.46% female). This is also comparable to results reported in the literature providing validity to the data if used for a screening population.[6] The positive predictive value was 58.47%; i.e. 58.47% of people with a positive screening test have a colorectal cancer/polyp diagnosis. The negative predictive value was 95.36%; i.e. 95.36% of people with a negative test result do not have a colorectal cancer/polyp diagnosis.

gFOBT result	Cancer /Polyp Diagnosis	Cancer/Polyp Diagnosis Free
Abnormal	533 colorectal cancers 993 polyps	1,084
Normal	396 colorectal cancers 967 polyps	28,031

Table 8: 2 by 2 table of colorectal cancer/polyp diagnosis by guaiac faecal occult blood test (gFOBT) result

3.3 Completeness of Records and Univariable Cox Regression

3.3.1 Completeness of Records for a Derived Screening Population

To determine the availability of data for a screening population which may be useful in a risk based prediction model, a cohort was derived of those aged 60-74 with a positive or negative result. **Table 9** summarises the frequency of these data and the completeness of variables. Since symptoms and diagnoses are binary parameters these are 100% recorded for individuals if they had consulted their GP. Registration details were also highly recorded, e.g. age, sex and GP practice (100%). Other more opportunistic factors such as anthropometrics were highly recorded. Lab measurements depending on the parameter were observed for around 45% of the screening population derived from the THIN

database (MCV, haemoglobin and platelet count). Ferritin was rarely recorded (8.59%). Although QOF indicators have been introduced for recording ethnic group, this factor was recorded in 54.76% of patient records. There is evidence to suggest that the ethnic group records are not representative of the UK population and so this parameter was not used for multivariable analysis but would be an important variable to consider in future models as it can capture certain genetic and social components.[50]

Lifestyle factors such as smoking status were extremely well recorded (99.44%) owing to the introduction of various QOF indicators for this parameter.[51] Alcohol consumption in units per week was reasonably well recorded at 78.00%. A summary of the continuous predictors and lab measurements for those with a cancer/polyp diagnosis and those without is given in **Table 10**. Boxplots for the lab measurements are also given in **Figure 3**. Ferritin is a highly skewed variable and would be better log transformed but since ferritin was rarely recorded (8.59%) it was dropped from further multivariable analysis.

Variable	Frequency	Total Observations	Missing	Percentage Recorded	Percentage Missing
Loss of Appetite					
Recorded	117				
Not Recorded	292051	292168	0	100.00	0.00
Combined Abdominal Pain and Antispasmodic Prescription					
Recorded	20797				
Not Recorded	271371	292168	0	100.00	0.00
Abdominal Pain					
Recorded	14209				
Not Recorded	277959	292168	0	100.00	0.00
Flatulence					
Recorded	498				
Not Recorded	291670	292168	0	100.00	0.00
Abdominal Mass					
Recorded	165				
Not Recorded	292003	292168	0	100.00	0.00
Antispasmodic drug prescription					
Recorded	9667				
Not Recorded	282501	292168	0	100.00	0.00
Anti-motility drug prescription					
Recorded	3616				
Not Recorded	288552	292168	0	100.00	0.00
Laxative Drug					
Recorded	23248				
Not Recorded	268920	292168	0	100.00	0.00

Abnormal Rectal Examination					
Recorded	3				
Not Recorded	292165	292168	0	100.00	0.00
Rectal Bleeding					
Recorded	2694				
Not Recorded	289474			100.00	0.00
Venous Thromboembolism					
Recorded	916				
Not Recorded	291252	292168	0	100.00	0.00
Tiredness					
Recorded	7176				
Not Recorded	284992	292168	0	100.00	0.00
Weight Loss					
Recorded	1057				
Not Recorded	291111	292168	0	100.00	0.00
Alcohol units per week (total observations)	227879				
Mean (SD)	9.50 (12.27)				
Min and Max	0-500	227879	64,289	78.00	22.00
Family History of Gastrointestinal Cancer					
Recorded	4424				
Not Recorded	287744	292168	0	100.00	0.00
Constipation					
Recorded	4263				
Not Recorded	287905	292168	0	100.00	0.00
Diarrhoea					
Recorded	5870				
Not Recorded	286298	292168	0	100.00	0.00
Sex					
Male	136569				
Female	155599	292168	0	100.00	0.00
Age at Latest FOBT (total observations)	292168				
Mean (SD)	66.43 (4.47)				
Min and Max	59-75	292168	0	100.00	0.00
Change in Bowel Habit					
Recorded	1656				
Not Recorded	290512	292168	0	100.00	0.00
BMI (total observations)	280032				
Mean (SD)	27.48 (5.01)				
Min and Max	11.4-60	280032	12,136	95.85	4.15
Height (total observations)	280670				
Mean (SD)	1.681 (0.10)				
Min and Max	1-2.41	280670	11,498	96.07	3.94
Weight (total observations)	282655				
Mean (SD)	78.08 (16.53)				

Min and Max	35-200	282655	9,513	96.74	3.26
Weight % change between two most recent readings (total observations)	90,761				
Mean (SD)	-0.07 (4.77)				
Min and Max	-56.01-281.8	90,761	201,407	31.07	68.94
BMI % change between two most recent readings (total observations)	90,516				
Mean (SD)	-0.076 (4.62)				
Min and Max	-55.95-170	90,516	201,652	30.98	69.02
Ferritin Continuous (total observations)	25090				
Mean (SD)	127.11 (201.74)				
Min and Max	1-10575	25090	267,078	8.59	91.41
Ferritin % change between two most recent readings (total observations)	10,290				
Mean (SD)	43.689 (334.26)				
Min and Max	-99,910 - 21900	10,290	281,878	3.52	96.48
Ferritin (**<15**µg/L vs >=15µg/L)					
<15µg/L	1252				
>=15µg/L	23838	25090	267,078	8.59	91.41
Mean Cell Volume Continuous	129,523				
	91.110 (5.08)				
	50.6-143.3	129,523	162,645	44.33	55.67
Mean Cell Volume % change between two most recent readings (total observations)	98,340				
Mean (SD)	0.066 (2.80)				
Min and Max	-35.56-79.350)	98,340	193,828	33.66	66.34
Mean Cell Volume (**<80fL** vs >=80fL)					
<80fL	2073				
>=80fL	127450	129523	162,645	44.33	55.67
Platelet Count Continuous (total observations)	129,728				
Mean (SD)	245.609 (66.00)				
Min and Max	1.76-1205	129,728	162,440	44.40	55.60
Platelet Count % change between two most recent readings (total observations)	98,603				
Mean (SD)	0.963 (40.23)				
Min and Max	-99.5-10606.32	98,603	193,565	33.75	66.25
Platelet Count (<=400x 109/L vs **>400 x 109/L**)					
>400 x 109/L	2764				
<=400x 109/L	126964	129728	162,440	44.40	55.60
Hb Continuous (total observations)	130039				
Mean (SD)	13.921 (1.30)				

Min and Max	2.86-23.8	130039	162,129	44.51	55.49
Hb % change between two most recent readings	98,968				
Mean (SD)	0.322 (6.60)				
Min and Max	-82.87-411.33	98,968	193,200	33.87	66.13
Hb (<11g/dL vs >=11g/dL)					
<11g/dL	1948				
>=11g/dL	128091	130039	162,129	44.51	55.49
Diabetes					
Recorded	32285				
Not Recorded	259883	292168	0	100.00	0.00
Crohn's disease					
Recorded	884				
Not Recorded	291284	292168	0	100.00	0.00
Ulcerative Colitis					
Recorded	1797				
Not Recorded	290371	292168	0	100.00	0.00
Irritable Bowel Syndrome					
Recorded	27112				
Not Recorded	265056	292168	0	100.00	0.00
Diverticulitis					
Recorded	18611				
Not Recorded	273557	292168	0	100.00	0.00
Previous Positive FOBTs					
0	290624				
1	1488				
2	54				
3	2	292168	0	100.00	0.00
Previous Negative FOBTs					
0	176584				
1	91111				
2	23478				
3	987				
4	8	292168	0	100.00	0.00
Previously screened with a FOBT					
No	175479				
Yes	116689	292168	0	100.00	0.00
Previous polyps diagnosed					
Recorded	7271				
Not Recorded	284897	292168	0	100.00	0.00
Primary care FOBT					
No	292136				
Yes	32	292168	0	100.00	0.00

Weight Loss (Percentage change since last two readings)					
<5 at baseline	82369				
5-9.9	6768				
=>10	1624	90761	201,407	31.07	68.94
Latest FOBT Result					
BCSP FOB test abnormal (baseline)	6362				
BCSP FOB test normal	285806	292168	0	100.00	0.00
Ethnic Group					
White (baseline)	151839				
Asian	4562				
Black	2048				
Mixed	610				
Other	931	159990	132,178	54.76	45.24
Urban Rural					
Town & Fringe – Less sparse (baseline)	38328				
Town & Fringe – Sparse	2726				
Urban >10k - Less sparse	220911				
Urban >10k – Sparse	152				
Village, Hamlet & Isolated	22440				
Village, Hamlet & Isolated	2136	286693	5,475	98.13	1.87
Townsend					
1 (baseline)	105661				
2	74316				
3	54656				
4	34592				
5 - most deprived	17407	286632	5,536	98.11	1.90
Blood Group					
A (baseline)	3117				
AB	250				
B	741				
O	3340	7448	284,720	2.55	97.45
Smoking Status					
Never smoked (baseline)	167941				
ex-smoker	97353				
current smoker	25244	290538	1,630	99.44	0.56

Table 9: The frequency and completeness in recording of investigated variables for an English colorectal cancer screening population who are adequately screened (positive/negative result).

	Patients with Colorectal Cancer/Polyp Diagnosis			Patients without Colorectal Cancer/Polyp Diagnosis		
Variable	**Mean**	**Standard Deviation**	**Frequency**	**Mean**	**Standard Deviation**	**Frequency**
Ferritin	109.45	141.70	336	127.35	202.43	24754
Haemoglobin Concentration	13.91	1.48	1544	13.92	1.30	128495
Mean Cell Volume	91.03	6.20	1537	91.11	5.07	127986
Weight (kg)	81.55	17.41	2805	78.04	16.52	279850
Height (m)	1.70	0.10	2794	1.68	0.10	277876
Platelet count	246.26	71.08	1540	245.60	65.93	128188
BMI	28.30	5.29	2786	27.48	5.01	277246
Cigarettes per day amongst smokers	12.48	7.76	228	11.84	7.49	16340
Age at FOBT	66.91	4.39	2,889	66.43	4.48	289,279
Alcohol units	11.82	14.79	2307	9.47	12.24	225572
Haemoglobin % change	0.19	7.83	1245	0.32	6.55	97723
Mean cell volume % change	-0.08	3.34	1237	0.07	2.79	97103
Ferritin % change	40.85	127.70	179	43.74	336.78	10111
BMI % change	-0.200	4.76	1175	-0.07	4.62	89341
Platelet count % change	0.79	16.55	1240	0.97	40.44	97363
Weight % change	-0.21	4.77	1,179	-0.07	4.78	89,582

Table 10: Summary of Continuous Predictors and Lab Measurements for those with a cancer/polyp diagnosis and those without.

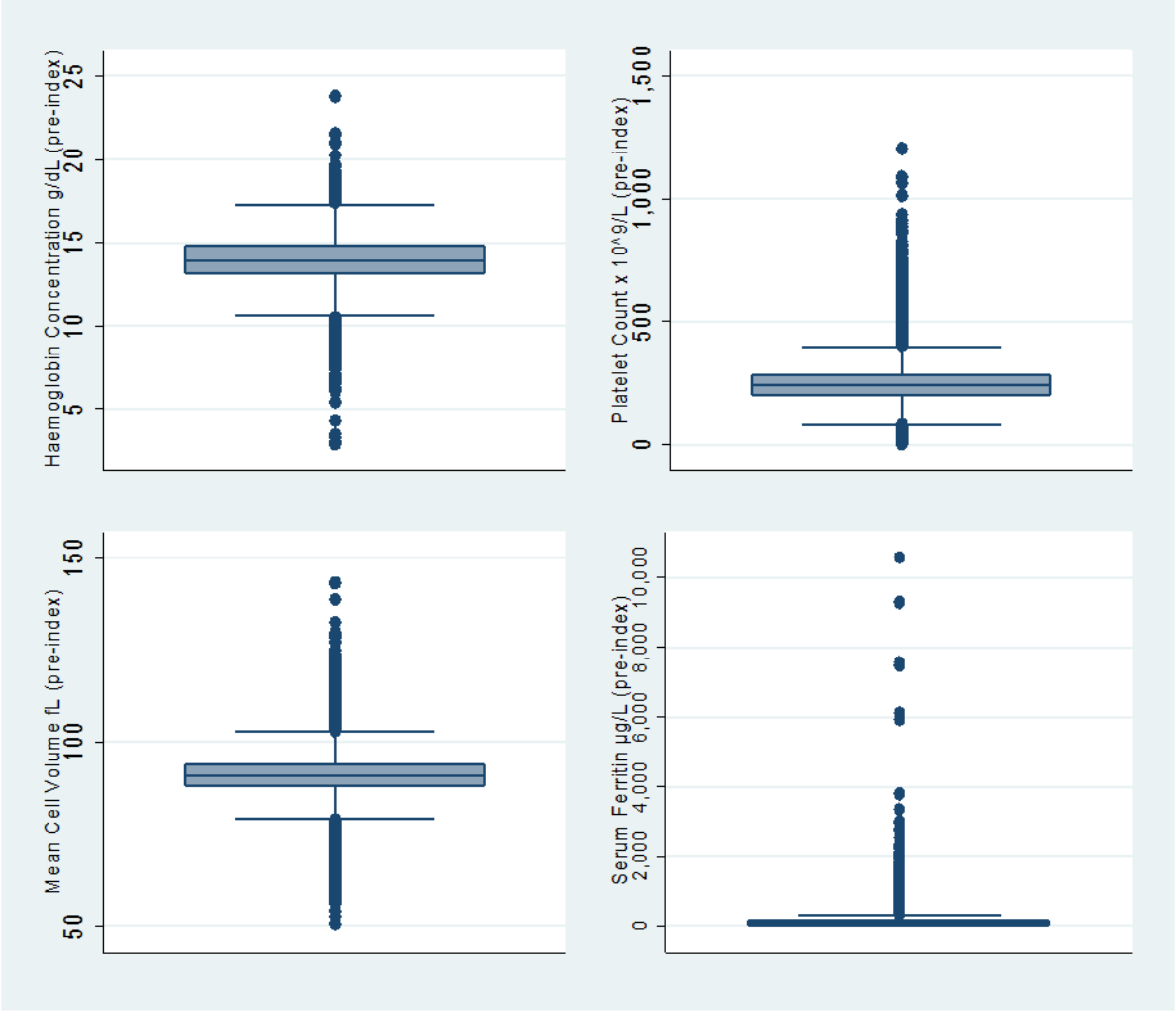


Figure 3: Boxplots for laboratory test results

3.3.2 Univariable Cox Regression

To determine the predictors with a strong independent association with colorectal cancer/polyps which could be included in a risk based prediction model, the univariable hazard ratios estimated using Cox Regression (with Efron ties) are presented for the variables of interest in **Table 11.**

Predictors with the largest observed hazard ratios included previous positive FOBT results with a hazard ratio (HR) of 5.032 (CI: 4.18-6.05), previous polyps diagnosed before the latest FOBT result (this could have been in or outside the screening programme) with a HR of 3.182 (CI: 2.768-3.659) and a symptom of rectal bleeding/melaena recorded with a HR of 3.118 (2.503-3.883). This means the cancer/polyp detection rate is 3 times as high for patients with previous polyps diagnosed or a rectal bleeding symptom and 5 times as high for those with previous positive FOBT results.

Continuous variables with a significant association included, age at FOBT (HR 1.025: CI 1.017-1.033), BMI (HR 1.029 CI: 1.022-1.036), height (HR 4.390: CI 3.001-6.421), weight (HR 1.012 CI 1.010-1.010), ferritin (HR 0.999 CI 0.069-0.998), and alcohol consumption in units per week (HR 1.010 CI: 1.008-1.011). This means that the hazard rate for the diagnosis of colorectal cancer/polyps would increase by 2.5% for each additional year in age at FOBT. In addition, females were at lower risk of colorectal cancer/polyp diagnosis than males (HR 0.656: CI 0.609-0.706).

Categorising lab measurements to define clinically relevant cutoffs led to haemoglobin, ferritin and MCV having a significant effect on the diagnosis of colorectal cancer/polyps with HRs of 2 and above. A platelet count of more than 400 x 10^9/L on the other hand which indicates thrombocytosis, did not have a significant HR (P = 0.379). Analysing lab results as continuous parameters led to non-significant results at a p value of 0.1.

Variable	Observed Hazard Ratio (estimated from the data)	Standard Error	z	P>z	95% Confidence Interval		Number of Observations
Loss of Appetite	2.614	1.510	1.66	0.096	0.842	8.108	292,059
Combined Abdominal Pain and Antispasmodic Prescription	1.424	0.089	5.67	0.000	1.261	1.610	292,059
Abdominal Pain	1.424	0.105	4.79	0.000	1.232	1.646	292,059
Flatulence	2.479	0.689	3.27	0.001	1.438	4.274	292,059
Abdominal Mass	1.259	0.890	0.33	0.745	0.315	5.035	292,059
Antispasmodic drug prescription	1.450	0.127	4.24	0.000	1.221	1.721	292,059
Anti-motility drug prescription	1.535	0.209	3.15	0.002	1.176	2.005	292,059
Laxative Drug	1.390	0.084	5.47	0.000	1.235	1.564	292,059
Rectal Bleeding	3.118	0.349	10.15	0.000	2.503	3.883	292,059
Venous Thromboembolism	1.421	0.395	1.26	0.206	0.824	2.451	292,059
Tiredness	1.358	0.141	2.95	0.003	1.108	1.664	292,059
Weight Loss	1.705	0.403	2.26	0.024	1.073	2.710	292,059
Alcohol units per week	1.010	0.001	9.93	0.000	1.008	1.011	227,792
Family History of Gastrointestinal Cancer	1.591	0.195	3.78	0.000	1.251	2.024	292,059
Constipation	1.654	0.200	4.16	0.000	1.305	2.097	292,059
Diarrhoea	1.778	0.177	5.79	0.000	1.463	2.160	292,059
Sex							292,059
Male	-	-	-	-	-	-	
Female	0.656	0.025	-11.23	0.000	0.609	0.706	
Age at Latest FOBT	1.025	0.004	5.91	0.000	1.017	1.033	292,059
Change in Bowel Habit	2.609	0.406	6.17	0.000	1.924	3.539	292,059
Weight % change between two most recent readings	0.994	0.006	-0.99	0.323	0.981	1.006	90,729
BMI	1.029	0.004	8.32	0.000	1.022	1.036	279,927
Height	4.390	0.852	7.62	0.000	3.001	6.421	280,563
Weight	1.012	0.001	11.05	0.000	1.010	1.01	282,550
BMI % change between two most recent readings	0.994	0.006	-0.87	0.385	0.982	1.007	90,484
Ferritin Continuous	0.999	0.000	-1.82	0.069	0.998	1.000	25,082
Ferritin % change between two most recent readings	1.000	0.000	-0.1	0.923	0.999	1.000	10,287
Ferritin (**<15**µg/L vs >=15µg/L)	2.054	0.377	3.93	0.000	1.434	2.942	25,082
Mean Cell Volume Continuous	0.996	0.005	-0.88	0.381	0.986	1.005	129,481
Mean Cell Volume % change between two most recent readings	0.981	0.010	-1.85	0.064	0.961	1.001	98,304
Mean Cell Volume (**<80fL** vs >=80fL)	2.419	0.326	6.54	0.000	1.856	3.151	129,481
Platelet Count Continuous	1.000	0.000	0.4	0.692	0.999	1.001	129,685
Platelet Count % change between two most recent readings	1.000	0.001	-0.12	0.901	0.998	1.002	98,566
Platelet Count (<=400x 109/L vs **>400 x 109/L**)	1.155	0.190	0.88	0.379	0.837	1.594	129,685
Hb Continuous	0.990	0.019	-0.52	0.605	0.953	1.029	129,996
Hb % change between two most recent readings	0.997	0.005	-0.76	0.449	0.988	1.006	98,931
Hb (**<11g/dL** vs >=11g/dL)	2.231	0.324	5.53	0.000	1.679	2.966	129,996
Diabetes	1.470	0.076	7.49	0.000	1.329	1.627	292,059
Crohn's disease	1.038	0.346	0.11	0.912	0.539	1.996	292,059

Ulcerative Colitis	1.686	0.309	2.85	0.004	1.177	2.416	292,059
Irritable Bowel Syndrome	1.141	0.069	2.17	0.030	1.013	1.286	292,059
Diverticulitis	1.226	0.086	2.92	0.004	1.069	1.406	292,059
Previous Positive FOBTs	5.032	0.474	17.16	0.000	4.184	6.052	292,059
Previous Negative FOBTs	0.770	0.026	-7.8	0.000	0.721	0.822	292,059
Previously screened with a FOBT	0.784	0.032	-6.01	0.000	0.724	0.849	292,059
Previous polyps diagnosed	3.182	0.227	16.26	0.000	2.768	3.659	292,059
Primary care FOBT	2.867	2.867	1.05	0.292	0.404	20.357	292,059
Weight Loss (Percentage change since last two readings)							90,729
<5 at baseline	-	-	-	-	-	-	
5-9.9	1.060	0.115	0.54	0.592	0.857	1.310	
=>10	0.948	0.214	-0.24	0.814	0.609	1.476	
Latest FOBT Result							292,059
BCSP FOB test normal (baseline)	-	-	-	-	-	-	
BCSP FOB test abnormal	55.849	2.086	107.72	0.000	51.908	60.090	
Ethnic Group							159,926
White (baseline)							
Asian	0.996	0.151	-0.02	0.980	0.741	1.340	
Black	0.786	0.197	-0.96	0.337	0.480	1.286	
Mixed	1.364	0.483	0.88	0.382	0.681	2.732	
Other	0.565	0.253	-1.28	0.202	0.235	1.359	
Urban Rural Town & Fringe – Less sparse (baseline)							286,591
Town & Fringe – Sparse	0.515	0.131	-2.6	0.009	0.313	0.849	
Urban >10k - Less sparse	0.876	0.046	-2.51	0.012	0.789	0.971	
Urban >10k – Sparse	1.741	1.008	0.96	0.339	0.559	5.418	
Village, Hamlet & Isolated	0.826	0.070	-2.26	0.024	0.700	0.975	
Village, Hamlet & Isolated	1.028	0.211	0.13	0.895	0.687	1.538	
Townsend							286,530
1 (baseline)							
2	1.065	0.052	1.28	0.199	0.968	1.172	
3	1.041	0.056	0.75	0.456	0.937	1.157	
4	1.161	0.070	2.47	0.014	1.031	1.308	
5 - most deprived	1.387	0.102	4.43	0.000	1.200	1.603	
Blood Group (A at baseline)							7,444
A (baseline)							
AB	0.442	0.450	-0.8	0.423	0.060	3.248	
B	0.708	0.343	-0.71	0.476	0.274	1.829	
O	0.975	0.254	-0.1	0.923	0.585	1.625	
Smoking Status							290,429
Non-Smoker (baseline)							
ex-smoker	1.532	0.061	10.72	0.000	1.417	1.656	
current smoker	1.618	0.098	7.91	0.000	1.436	1.823	

Table 11: Univariable Cox Regression for considered variables with associated hazard ratios.

Fractional Polynomials for Continuous Variables

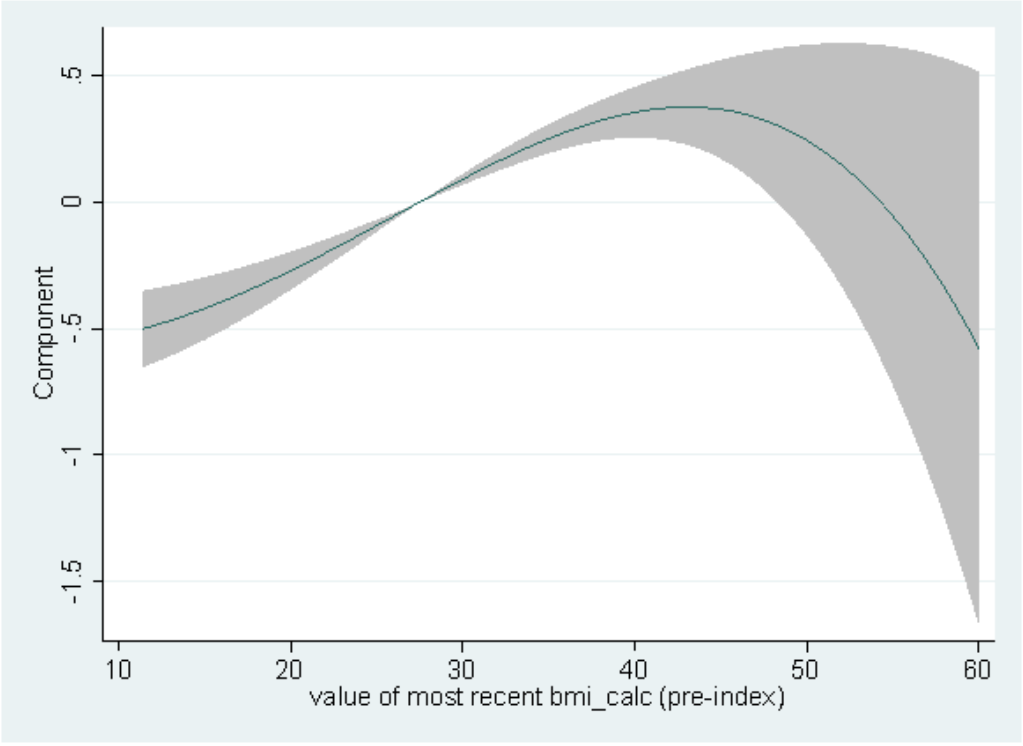
Fractional polynomials can be used to provide flexible parameterisation for continuous variables and to determine the most appropriate functional form of a covariate in the model. This modelling method was investigated for; BMI, alcohol consumption in units per week, haemoglobin concentration, mean cell volume and platelet count. The most efficient model was selected using the likelihood ratio test at a p value of 0.1.

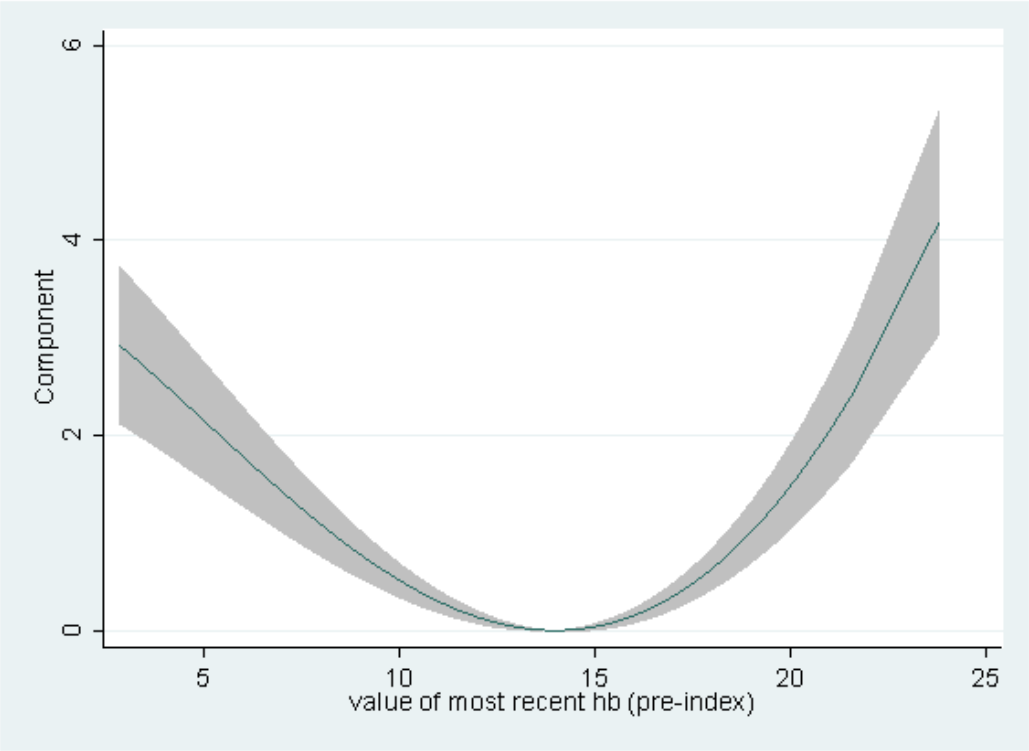
For age at FOBT, the linear model and fractional polynomials of degree 1 (FP1) were the best fitting, using a 0.1 significance level the linear model is selected for model simplicity. For BMI, a fractional polynomial of degree 2 (FP2) has the best fit to these data ($\beta_1(BMI)^3$ + $\beta_2(BMI)^3\ln(BMI)$). Alcohol units also have the best fit using an FP2 model ($\beta_1(alcohol)^{-0.5}$ + $\beta_2\ln(alcohol)$) as well as Hb concentration ($\beta_1(Hb)^2$ + $\beta_2(Hb)^2\ln(Hb)$), mean cell volume ($\beta_1(MCV)^{0.5}$ + $\beta_2(MCV)^{0.5}\ln(MCV)$) and platelet count ($\beta_1(platelet)^{0.5}$ + $\beta_2(platelet)^{0.5}\ln(platelet)$) (**Table 12**). Any variable to the power zero is the natural log using the 'fp' function in Stata. Fractional polynomial component plots which show the fit of the model/fractional polynomial for each variable are shown in **Figure 4** with the 95% confidence interval.

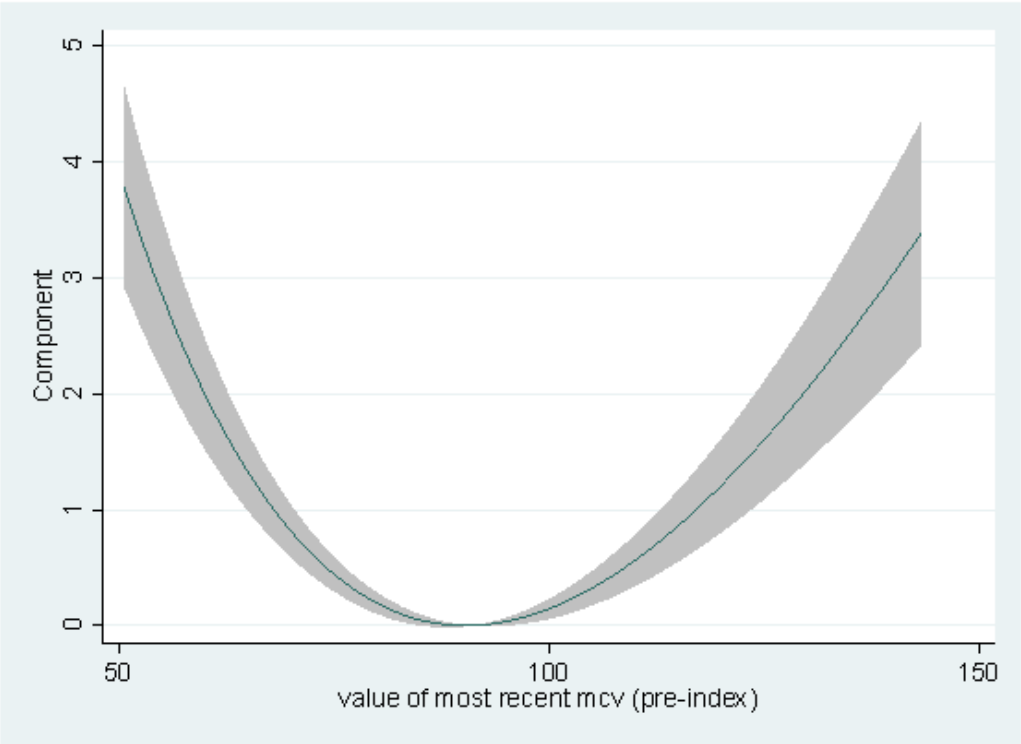
Multivariable fractional polynomials were considered when developing the multivariable Cox models.

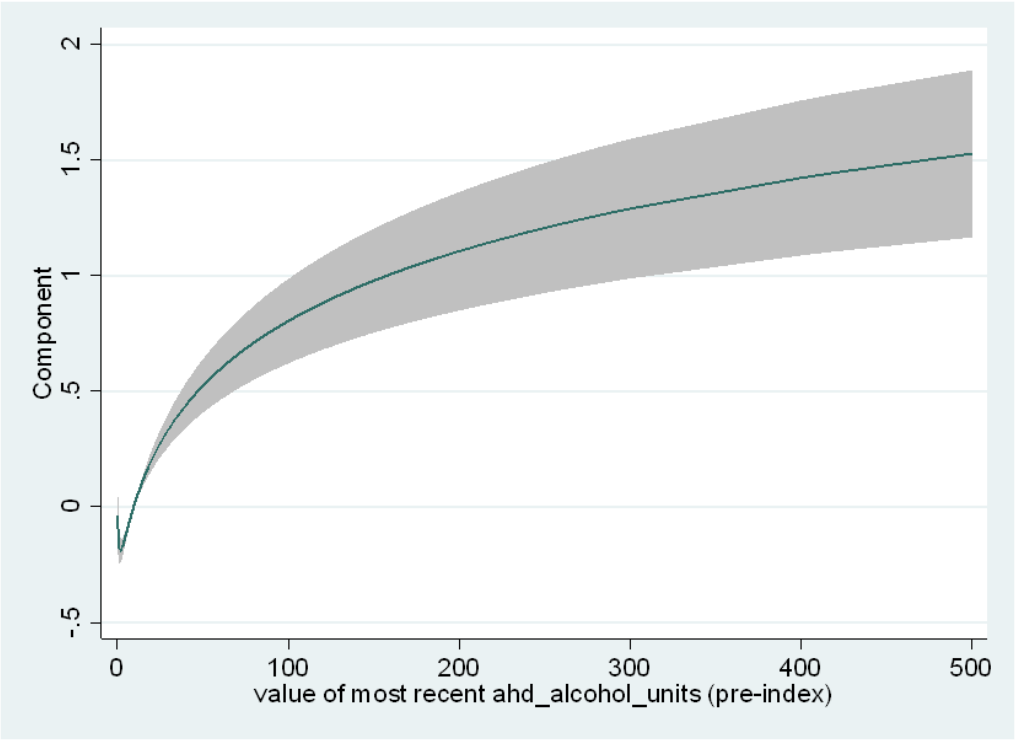
Predictor	Hazard Ratio	Standard Error	z	P>z	95% CI	
$\beta_1(BMI)^3$	1.067	0.013	5.170	0.000	1.041	1.094
$\beta_2(BMI)^3\ln(BMI)$	0.964	0.008	-4.490	0.000	0.949	0.980
$\beta_1(alcohol)^{-0.5}$	1.181	0.036	5.530	0.000	1.114	1.253
$\beta_2\ln(alcohol)$	1.663	0.112	7.590	0.000	1.459	1.897
$\beta_1(Hb)^2$	0.049	0.020	-7.340	0.000	0.022	0.109
$\beta_2(Hb)^2\ln(Hb)$	37.415	18.060	7.500	0.000	14.527	96.365
$\beta_1(MCV)^{0.5}$	7.67E+07	1.67E+08	-8.370	0.000	1061691	5.54E+09
$\beta_2(MCV)^{0.5}\ln(MCV)$	0.135	0.098	8.310	0.000	0.033	0.560
$\beta_1(platelet)^{0.5}$	0.135	0.098	-2.760	0.006	0.033	0.560
$\beta_2(platelet)^{0.5}\ln(platelet)$	29.252	31.229	3.160	0.002	3.609	237.085

Table 12: Fractional polynomials for investigated continuous variables. Any variables to the power 0 is the natural log using the 'fp' function in Stata.


Component
.5
0
-.5
-1
-1.5
10
20
30
40
50
60
value of most recent bmi_calc (pre-index)


Component
6
4
2
0
5
10
15
20
25
value of most recent hb (pre-index)


Component
0
1
2
3
4
5
50
100
150
value of most recent mcv (pre-index)


Component
-.5
0
.5
1
1.5
2
0
100
200
300
400
500
value of most recent ahd_alcohol_units (pre-index)

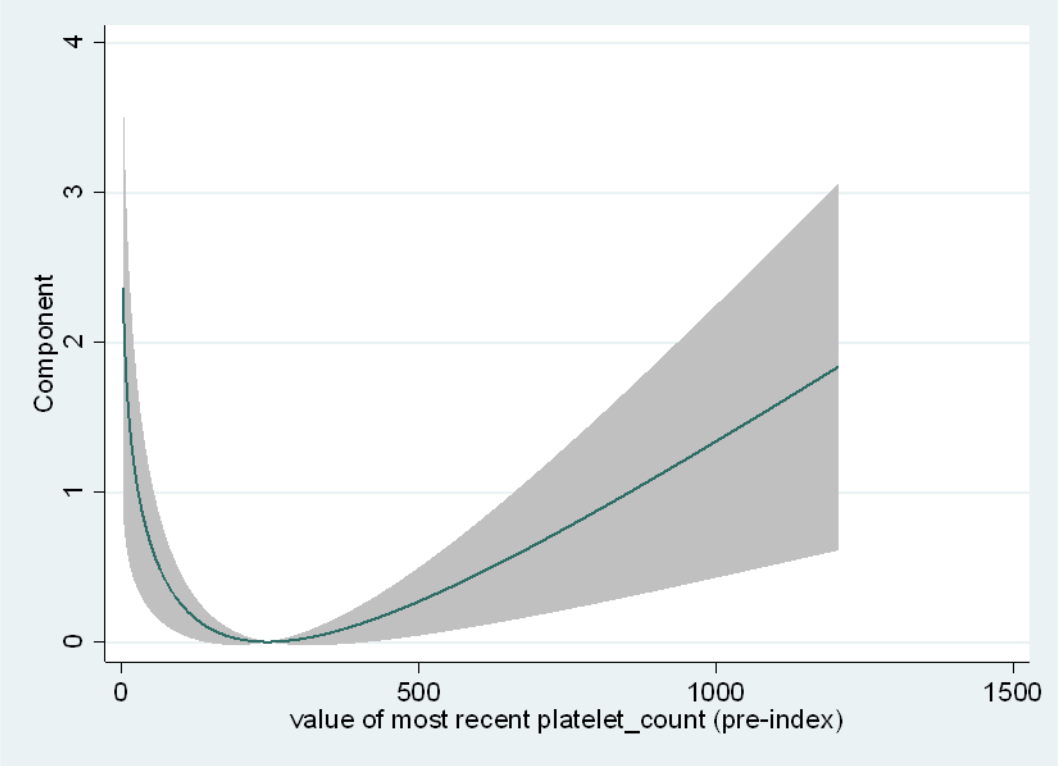


Figure 4: Fractional Polynomial component-plus-residuals plots with 95% confidence Intervals for; BMI (1st plot), Hb concentration (2nd plot), MCV (3rd plot), alcohol units per week (4th plot), platelet count (5th plot). The shaded region is the 95% confidence interval.

3.4 Kaplan Meier Survival Curve Analysis

This section helped to describe and check the validity of the data extracted for analysis and identified predictors which affect survival by comparing the survival functions of different groups/covariate patterns.

3.4.1 Survival Analysis - Time to Diagnosis (Colorectal Cancer Free Survival)

For survival analysis using the screening cohort there were 292,059 observations and 2889 events (113,454,249 person years). The median time to diagnosis ($S(t_{0.5})$) (or median survival time) after a FOBT and censoring at 2 years follow up was 383 days (see boxplot **Figure 5**).

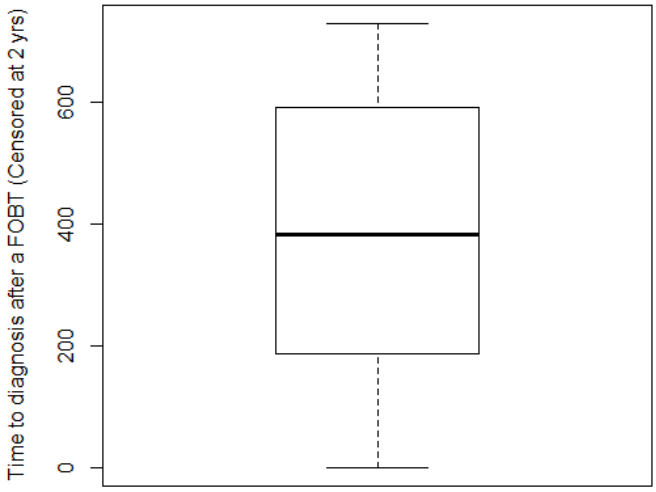
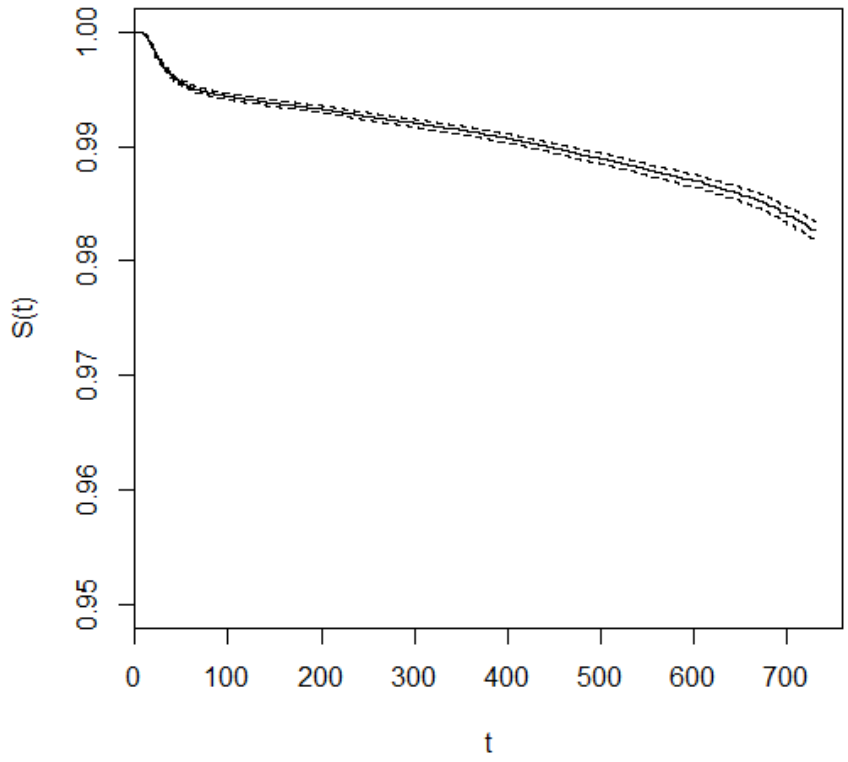


Figure 5: Boxplot of time to diagnosis after the index date (latest FOBT) when censoring data at 2 year follow up for the derived screening cohort.

The Kaplan-Meier estimate of the survivor function ($S(t)$) is a nonparametric maximum likelihood estimate (MLE) and is shown in **Figure 6** with the corresponding risk table. The shallow decline for the first year shows that there are reasonably low numbers of individuals diagnosed. There is a slightly steeper gradient for year two with more individuals being diagnosed.



Time (days)	Number at Risk	Number of Events	Survival	Standard Error	Lower 95% CI	Upper 95% CI
0	292168	0	1.000	0.000000	1.000	1.000
100	253654	1588	0.994	0.000141	0.994	0.995
200	214230	258	0.993	0.000157	0.993	0.994
300	177810	243	0.992	0.000175	0.992	0.992
400	140016	207	0.991	0.000197	0.990	0.991
500	104172	216	0.989	0.000230	0.989	0.989
600	70418	173	0.987	0.000274	0.986	0.988
700	38789	161	0.984	0.000368	0.983	0.985

Figure 6: Kaplan Meier estimate of the survivor function for time to diagnosis (colorectal cancer free survival) for the derived screening cohort with the corresponding risk table.

A Kaplan Meier estimate (using 2 year censoring for follow up) is plotted for those with a negative FOBT and those with a positive FOBT in **Figure 7** to determine the difference in cancer free survival between these groups of patients. The log-rank test for equality of survivor functions is significant between the two results ($p < 0.001$). This test assesses if there is a statistical difference between the survival times between those with a positive and those with a negative FOBT. From the plot, as expected, those with a positive FOBT receive more diagnoses over time. Initially with the sharp decline in cancer free survival (increase in diagnosis) associated with follow up after a positive screening result by the BCSP.

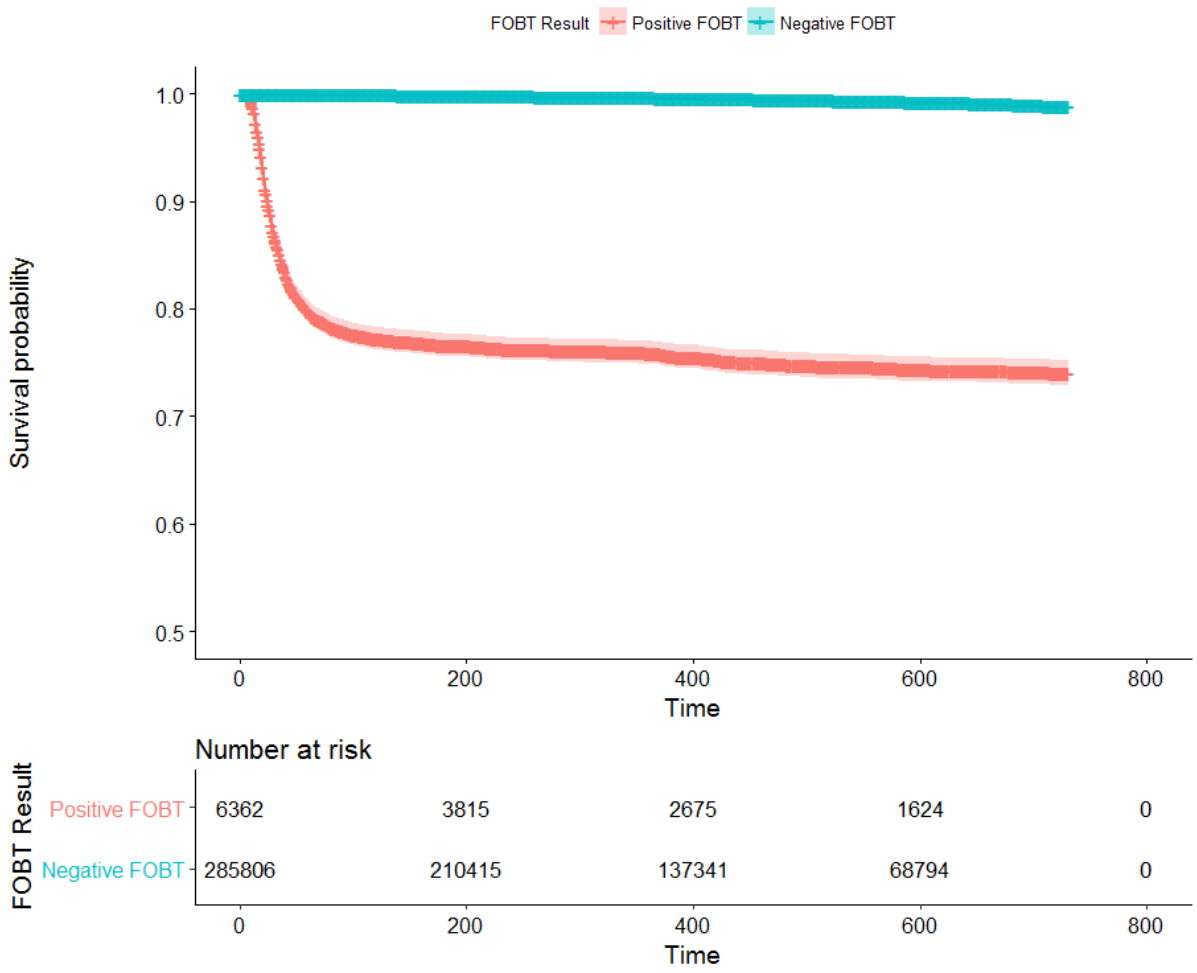


Figure 7: Kaplan Meier estimate for colorectal cancer free survival using 2 year censoring for the screening cohort plotted by negative or positive FOBT. Time is in days. The associated risk table is also displayed below the plot.

The Kaplan Meier estimates are plotted by sex in **Figure 8** with a significant log-rank test ($p<0.001$). Males have a significant reduction in colorectal cancer free survival (increase in colorectal cancer/polyp diagnoses) over time compared to females for the two year follow up period with the difference becoming more apparent over time. This is most likely due to the increased risk of CRC and increased FOBT positivity seen in men versus women.[77]

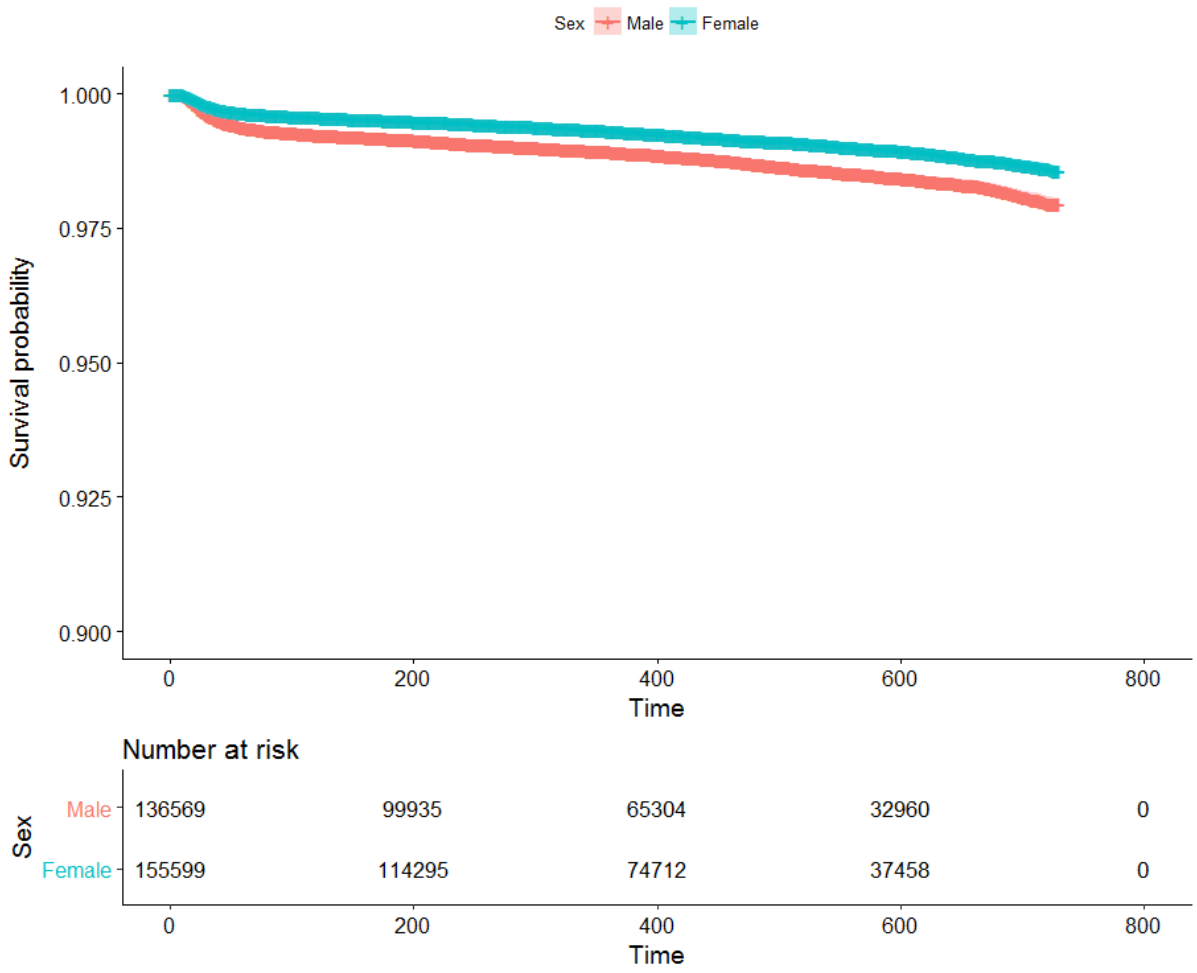


Figure 8: Kaplan Meier estimates for colorectal cancer free survival plotted by sex for the derived screening cohort with associated risk table below the plot.

3.4.2 Survival Analysis - Time to Death (Overall Survival)

Time to death was investigated for the screening cohort covering the period 1st May 2009 to 17th January 2017. The median time to death (from any cause) after a FOBT ($S(t_{0.5})$) (or median survival time) is 387 days (**Figure 9**). The Kaplan-Meier estimate of the survivor function ($S(t)$) for time to death is presented in **Figure 10** below with the corresponding risk table.

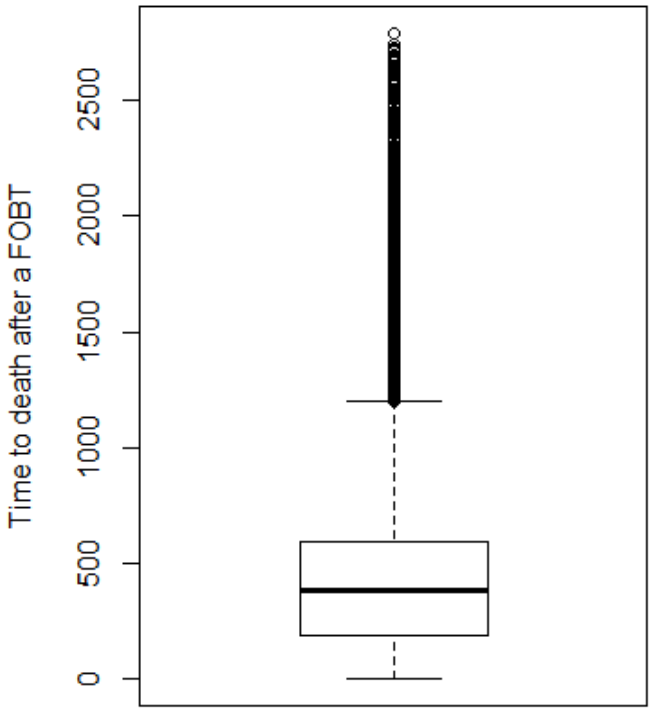
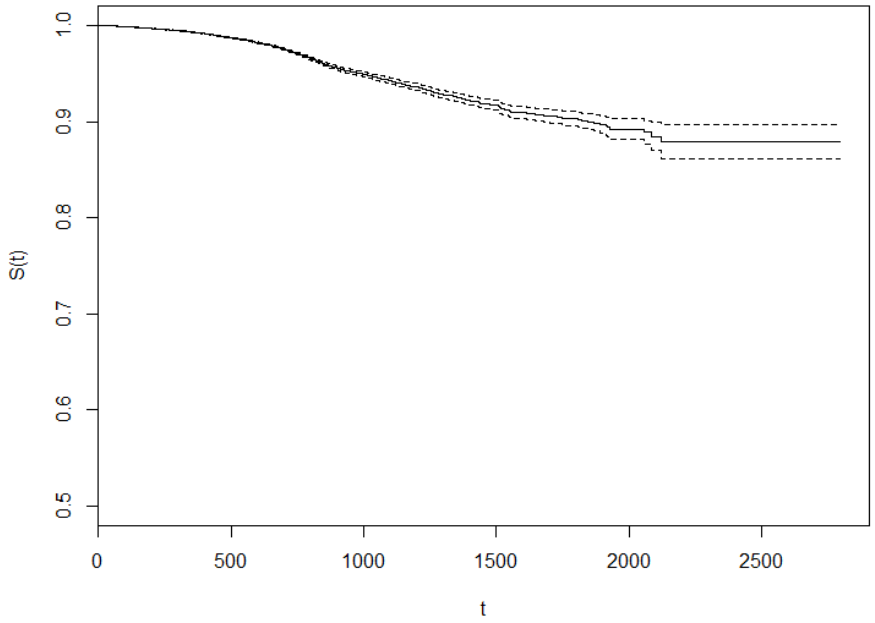


Figure 9: Boxplot of time to death after the index date (latest FOBT) for the derived screening cohort covering the period 1st May 2009 to 17th January 2017.



Time (days)	Number at Risk	Number of Events	Survival	Standard Error	Lower 95% CI	Upper 95% CI
0	292168	2	1.000	4.84e-06	1.000	1.000
100	255139	327	0.999	6.67e-05	0.999	0.999
200	215792	468	0.997	1.14e-04	0.997	0.997
300	179400	409	0.995	1.53e-04	0.994	0.995
400	141521	504	0.992	2.07e-04	0.991	0.992
500	105584	522	0.987	2.77e-04	0.987	0.988
600	71706	475	0.982	3.69e-04	0.981	0.983
700	39892	366	0.975	5.08e-04	0.974	0.976
800	18720	268	0.966	7.89e-04	0.964	0.967
900	13681	166	0.956	1.10e-03	0.953	0.958
1000	10575	80	0.949	1.30e-03	0.947	0.952
1100	8484	69	0.942	1.53e-03	0.939	0.945
1200	6598	52	0.936	1.76e-03	0.933	0.939
1300	5006	50	0.928	2.08e-03	0.924	0.932
1400	3636	28	0.922	2.37e-03	0.917	0.927
1500	2554	17	0.917	2.66e-03	0.912	0.922
1600	1771	18	0.910	3.14e-03	0.903	0.916
1700	1305	7	0.905	3.51e-03	0.899	0.912
1800	909	3	0.903	3.75e-03	0.896	0.910
1900	594	5	0.897	4.59e-03	0.888	0.906
2000	352	3	0.892	5.40e-03	0.882	0.903
2100	198	2	0.885	7.54e-03	0.870	0.899
2200	97	1	0.879	9.18e-03	0.861	0.897
2300	60	0	0.879	9.18e-03	0.861	0.897
2400	41	0	0.879	9.18e-03	0.861	0.897
2500	28	0	0.879	9.18e-03	0.861	0.897
2600	16	0	0.879	9.18e-03	0.861	0.897
2700	6	0	0.879	9.18e-03	0.861	0.897

Figure 10: Kaplan Meier estimate for time to death for the derived screening population with associated risk table using the cohort covering the period 1st May 2009 to 17th January 2017.

The Kaplan Meier estimates for time to death (in days) are plotted by FOBT result in **Figure 11.** Those with a positive result are at increased risk of death for around the first 1250 days, the risk then becomes similar and then for the last 1000 days those with negative FOBT results then appear to be at greatest risk. Presumably because cancers may have been missed (false negatives) or aggressive faster growing cancers have developed over the time period (interval cancers) which were missed/not present at screening. An alternative explanation may be because the positive FOBT group has seen quite a few 'sicker' people die due to CRC and thus after a certain time this makes them an overall 'healthy' group compared to the negatives.

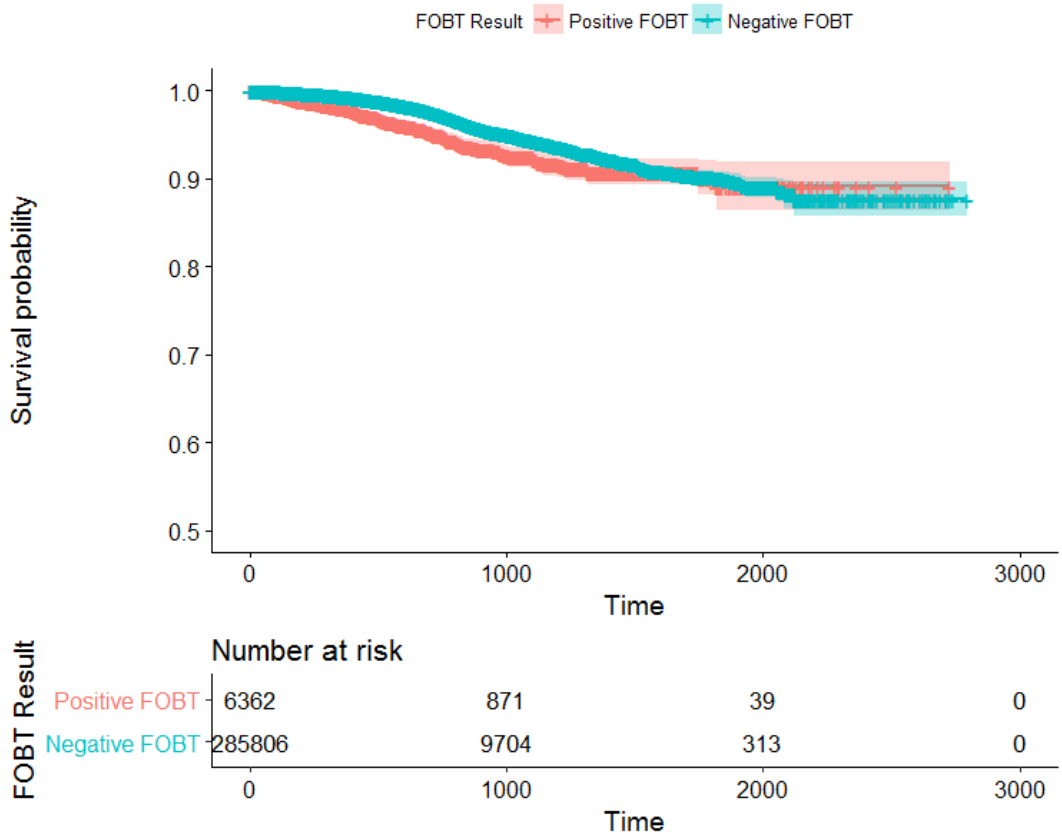


Figure 11: Kaplan Meier estimate for time to death for the derived screening cohort plotted by negative or positive FOBT. The associated risk table is also displayed below the plot.

3.4.3 Subgroup analysis negative FOBT

For time to diagnosis (colorectal cancer free survival) and for a sample population of the screening cohort who just had negative FOBT results, there were 285,697 observations and 1363 events (111,277,576 person years). The Kaplan-Meier survival estimate is given in **Figure 12** below with the corresponding risk table.

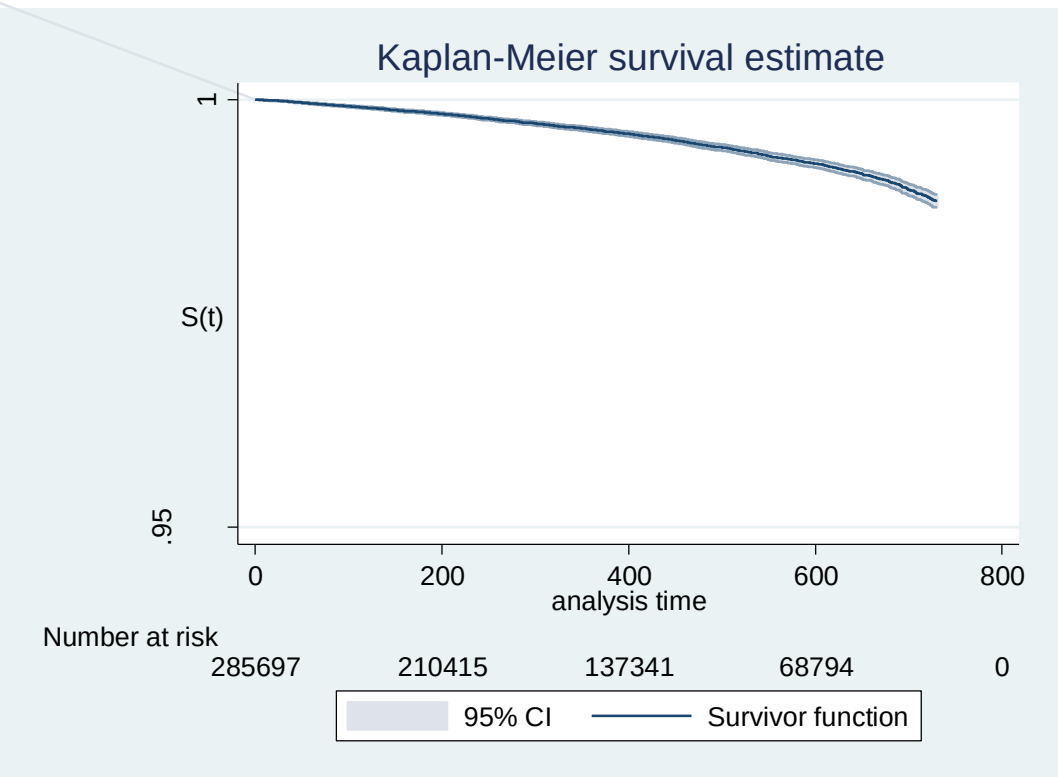


Figure 12: Kaplan Meier estimate of time to diagnosis (colorectal cancer free survival) for those with negative FOBT results censored at 2 years of follow up.

The Kaplan Meier estimates are plotted by sex in **Figure 13.** The log rank test of equality across strata is significant between males and females $p<0.001$ with FOBT negative males having a greater risk of diagnosis over time (a two year screening round) compared to FOBT negative females. This is as expected since males are in general at greater risk of developing colorectal cancer compared to females.

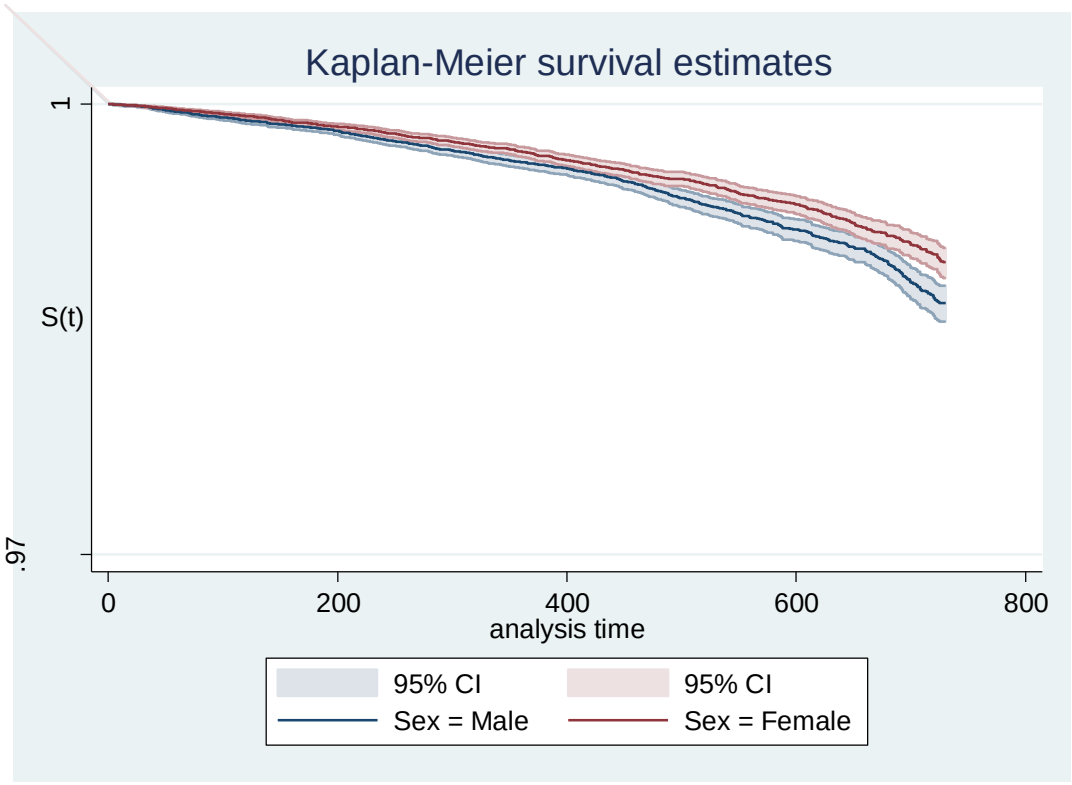


Figure 13: Kaplan Meier estimates of time to diagnosis (colorectal cancer free survival) plotted by sex for those with negative FOBT results censored at 2 years of follow up.

Time to death (overall survival) was also analysed for this population using the Kaplan-Meier survival estimate (**Figure 14**). For the population with just negative FOBT results there were 285,696 observations, 3606 failures and 119,018,674 person years. The last observed exit was at 2794 days.

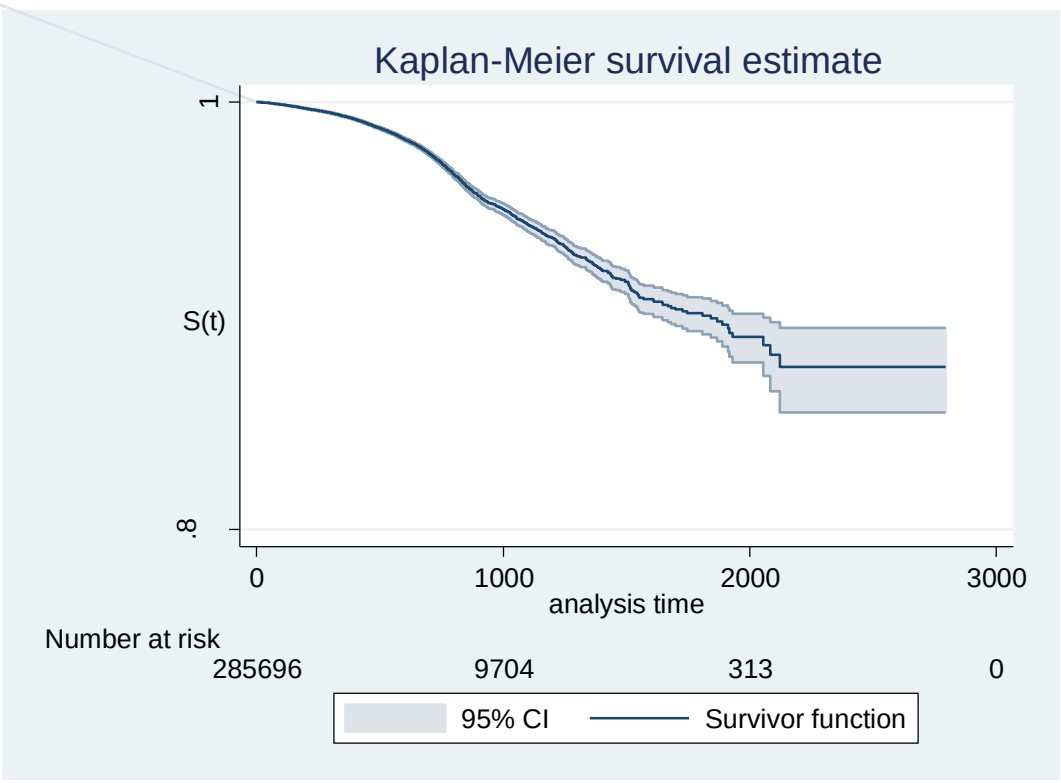


Figure 14: Kaplan Meier estimate of time to death for those with negative FOBT results.

The Kaplan-Meier survival estimate is also presented by sex below (**Figure 15**) with a significant log-rank test for males with negative FOBT results versus females with negative FOBT results p<0.001. Males are at greater risk of death over the study period.

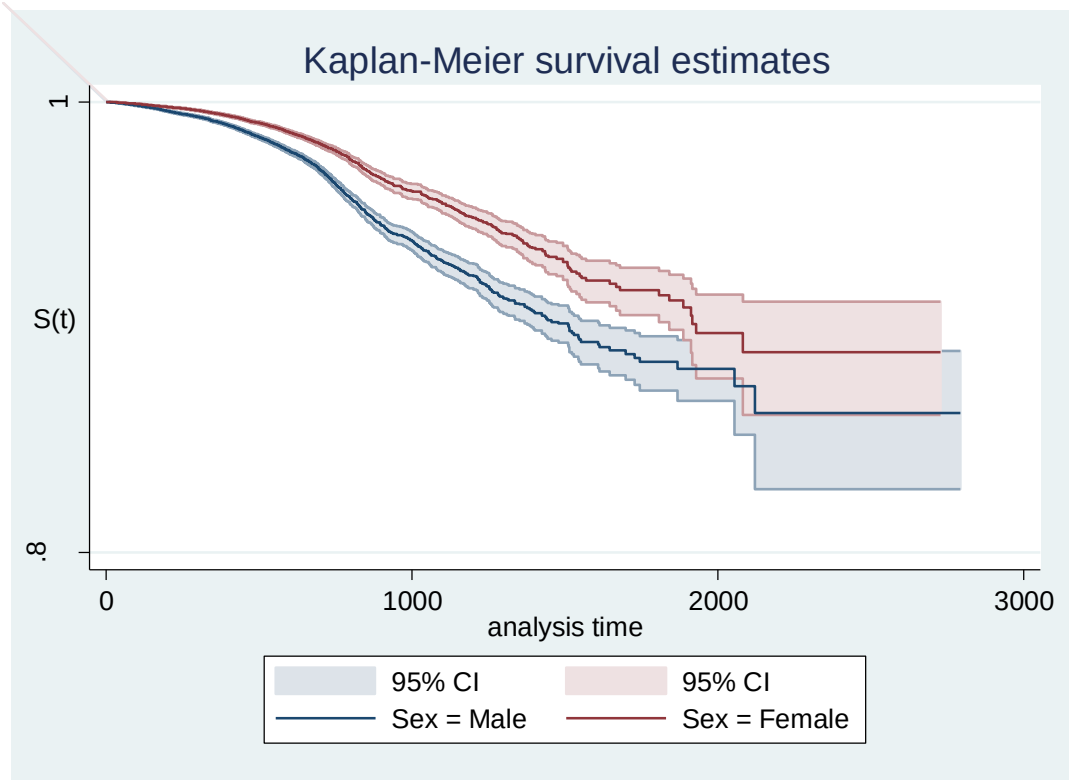


Figure 15: Kaplan Meier estimate of time to death for those with negative FOBT results by sex. This had a significant log rank test p=<0.0001.

3.4.4 Subgroup analysis TP, TN, FP, FN

Time to Diagnosis (Colorectal Cancer Free Survival)

Kaplan-Meier estimates for true positive results (TP), true negatives (TN), false positives (FP) and false negatives (FN) are presented in **Figure 16** below. This used a restricted dataset to ensure that everyone without a diagnosis had at least 2 years of follow up (n=32,004). The difference between the survival times between these groups is significant using a log-rank test (p<0.0001). FPs and TNs have a similar survival probability pattern. FNs have a constant rate of decline in survival (or increase in cancer diagnosis) over the study period. TPs have the steepest decline in survival probability as expected.

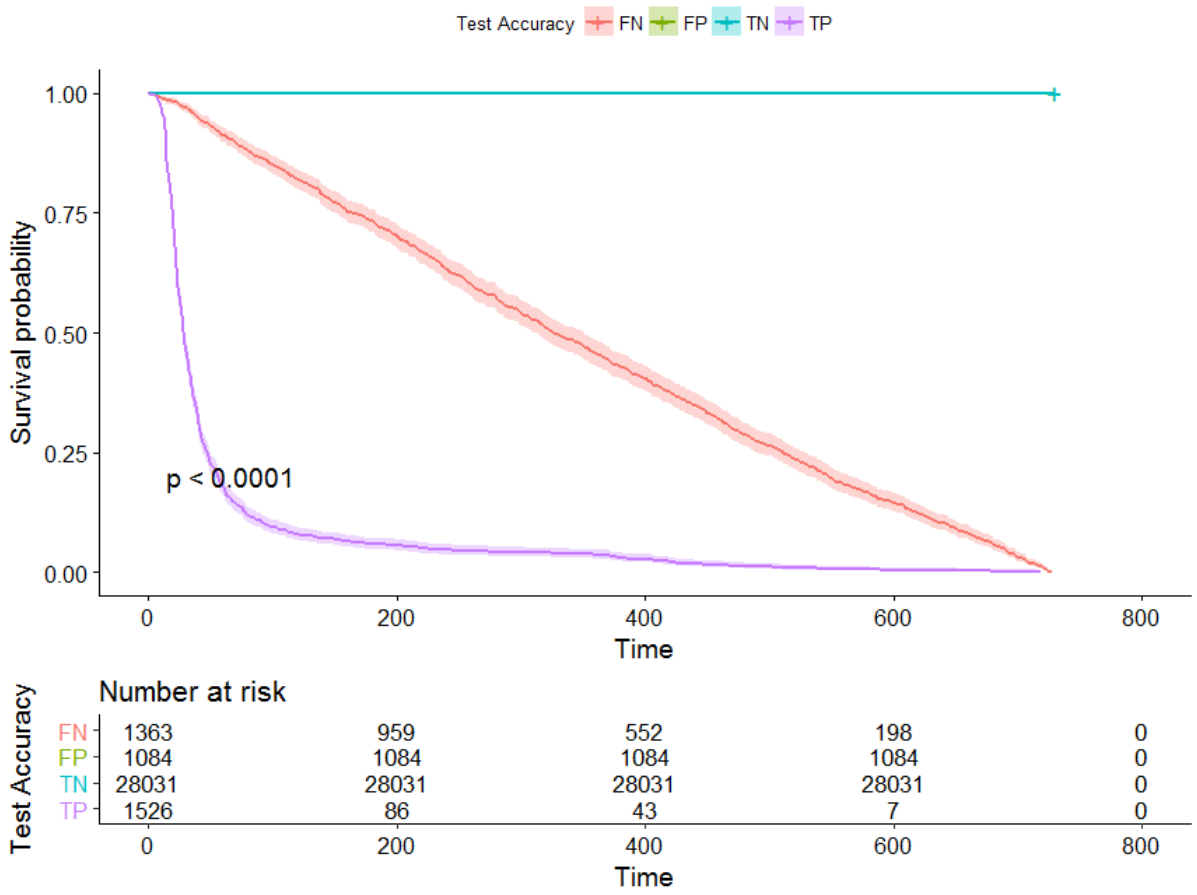


Figure 16: Kaplan-Meier estimates for time to diagnosis (colorectal cancer free survival) true positive results (TP), true negatives (TN), false positives (FP) and false negatives (FN) in the sample population with at least 2 years follow up if undiagnosed. The associated risk table is presented below.

Time to Death (Overall Survival)

Kaplan Meier estimates using time to death as the outcome are presented for FNs, FPs, TNs and TPs below (**Figure 17**). The FNs end up doing worse later on than the TPs at around 1000 days, this could be due to the effects of late diagnosis for FNs as they are less likely to have had diagnostic follow up/further investigations compared to TPs. In addition, FNs could have more aggressive cancers that develop after screening and TPs present at an earlier stage. FPs also appear to have fewer diagnoses than TNs possibly due to having some sort of diagnostic investigation putting them at lower risk of CRC compared to TNs.

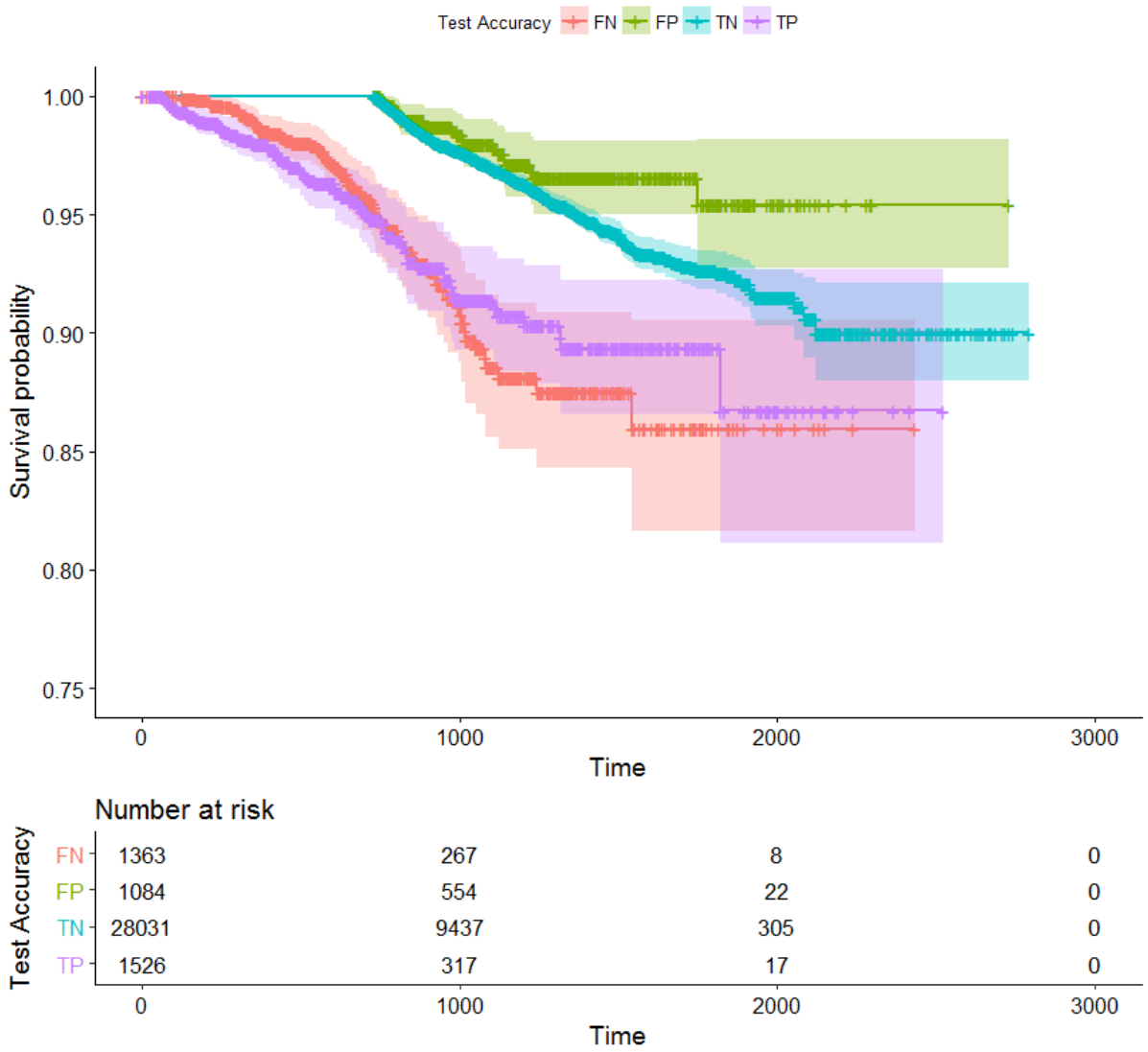


Figure 17: Kaplan-Meier estimates for time to death for true positive results (TP), true negatives (TN), false positives (FP) and false negatives (FN) in the sample population with at least 2 years follow up if undiagnosed. The associated risk table is presented below.

3.5 Multivariable Analysis Risk Prediction Model Development (n=98,303)

3.5.1 Cox Regression for Positive and Negative Results

This section identified additional predictors from a multivariable model which have the potential to be added to a risk based model to make more accurate screening referral decisions.

Survival models were built using Cox regression with backwards elimination and likelihood ratio testing for model selection, at a p-value of 0.05 using the 'mfp' function implemented in Stata.[55] There were 98,303 patients, 1,197 failures and 38,005,604 total analysis time at risk. There were 2511 positive test results and 95,792 negative FOBT results. The median follow up time in days was 385 (95% CI: 383 to 388), the restricted mean survival time was 722.181 (the largest observed analysis time was censored so the mean is underestimated). Original model had 33 degrees of freedom with all predictors considered, 45 degrees of freedom when including all considered interactions.

The final model after assessing all eligible predictors included; FOBT result, smoking status, whether a patient had a diagnosis of Crohn's disease, previous polyps diagnosed, flatulence, MCV of <80fL compared to a MCV of ≥80fL, alcohol consumption in units per week, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit (**Table 13**). Significant interactions at the 0.05 p value level included FOBT result and age and MCV and age. The adjusted hazard ratios for this model are reported in **Appendix 6**.

Multivariable fractional polynomials were used when model building to assess continuous predictors, both alcohol and age were centred around the mean with the natural log being used for alcohol consumption as this better fitted linearity. For this model, Harrell's C statistic was 0.854 (95% CI: 0.841, 0.868) and Somers D 0.708. Harrell's C statistic means that the predictors used in the model correctly identify the order of survival times for pairs of patients 85% of the time. The final model had 16 degrees of freedom with an AIC of 23499.87 and BIC 23581.27 (N=1197 when calculating BIC)). Overall model fit was assessed using adjusted R^2 which was 0.563 (bootstrapped CI 100 reps: 0.535, 0.596) and adjusted D was 2.321.[73 78] Regular R^2 was 0.568 with D statistic of 2.344.

Interactions which were significant at the 0.05 p-value level were FOBT result and age at FOBT, MCV <80fL and age at FOBT. The model with the interaction was compared to the

model without the interaction using the likelihood ratio test. This was statistically significant at the 0.05 level which indicates that the interactions improve the fit of the model to the data (Prob > chi^2 = 0.0019).

Variable	Observed Coefficient	Bootstrapped Standard Error	z	P>z	[95% Confidence Intervals]	
MCV*age at FOBT interaction	0.086	0.042	2.04	0.041	0.004	0.169
FOBT result*age at FOBT interaction	-0.037	0.014	-2.63	0.009	-0.065	-0.009
FOBT Result (positive)	3.741	0.061	61.61	0.000	3.622	3.860
Smoking Status:						
ex-smoker	0.206	0.060	3.44	0.001	0.089	0.324
current smoker	0.323	0.105	3.07	0.002	0.117	0.530
Crohn's Disease Diagnosis Recorded	-0.722	0.425	-1.7	0.089	-1.555	0.110
Previous Polyps Diagnosed	0.648	0.141	4.6	0.000	0.372	0.924
Flatulence Symptom Recorded	0.850	0.433	1.96	0.050	0.000	1.700
MCV <80fL	0.344	0.201	1.71	0.087	-0.050	0.738
Alcohol consumption (units per week)	0.082	0.031	2.67	0.008	0.022	0.142
Family History of Gastrointestinal Cancer	0.766	0.181	4.23	0.000	0.411	1.121
Abdominal pain/antispasmodic prescription recorded	0.199	0.098	2.02	0.043	0.006	0.392
Diarrhoea symptom	0.272	0.155	1.76	0.079	-0.031	0.575
Sex	-0.196	0.072	-2.72	0.007	-0.337	-0.055
Age at FOBT	0.033	0.008	3.85	0.000	0.016	0.049
Change in bowel habit symptom	0.908	0.200	4.55	0.000	0.517	1.300

Table 13: Cox regression model (coefficients) after mfp selection for patients with a positive or negative FOBT. The continuous variable age at FOBT has been centred (age_at_FOBT-66.97), alcohol units have the following transformation (ln(X)+2.25: X = (ahd_alcohol_units+1)/100). The deviance of the model is 23,467.87.

The linear predictor from the final model had a mean of 0.135 and a standard deviation of 0.688 (range: -1.372 to 5.758). The distribution of the linear predictor is shown in **Figure 18**. The linear predictor is the linear combination of predictors in the model without the baseline hazard. Discrimination was also assessed by analysing the separation between Kaplan-Meier curves for 4 risk groups (where the linear predictor is divided into 4 groups) (see **Figure 19**). Separation is much greater for group four compared to the other groups which most likely reflects a positive FOBT result and quick referral pathway (shorter time to diagnosis compared to other groups).

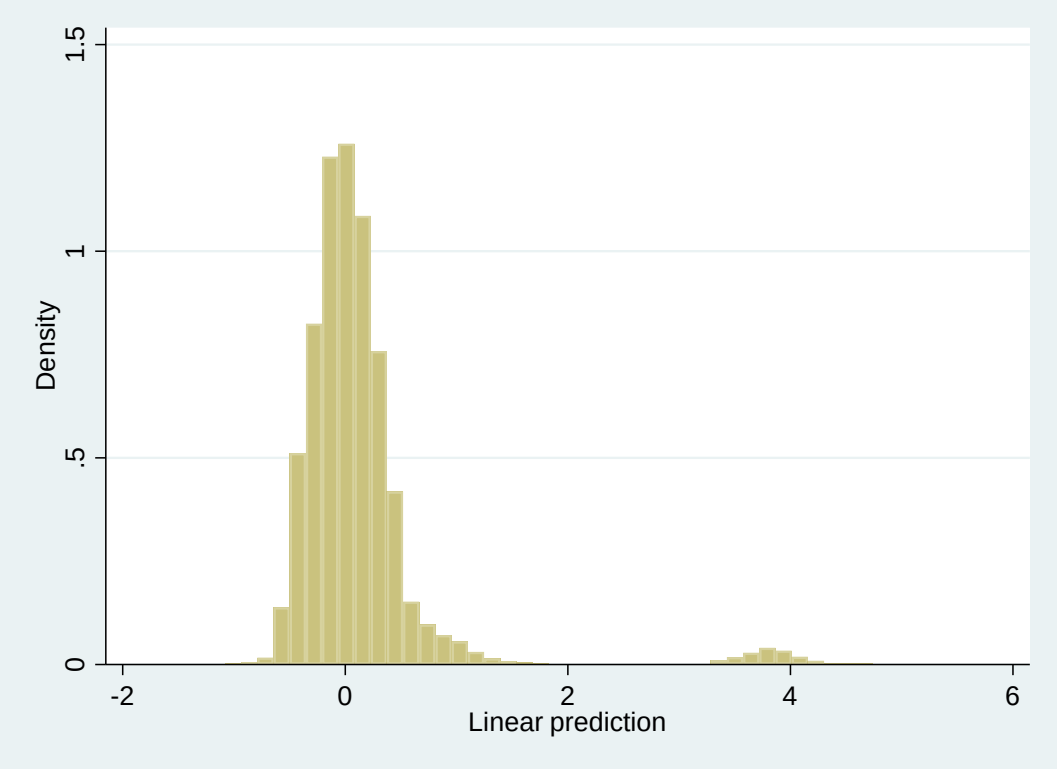


Figure 18: Distribution of the linear predictor for the final multivariable model for patients with positive and negative FOBTs

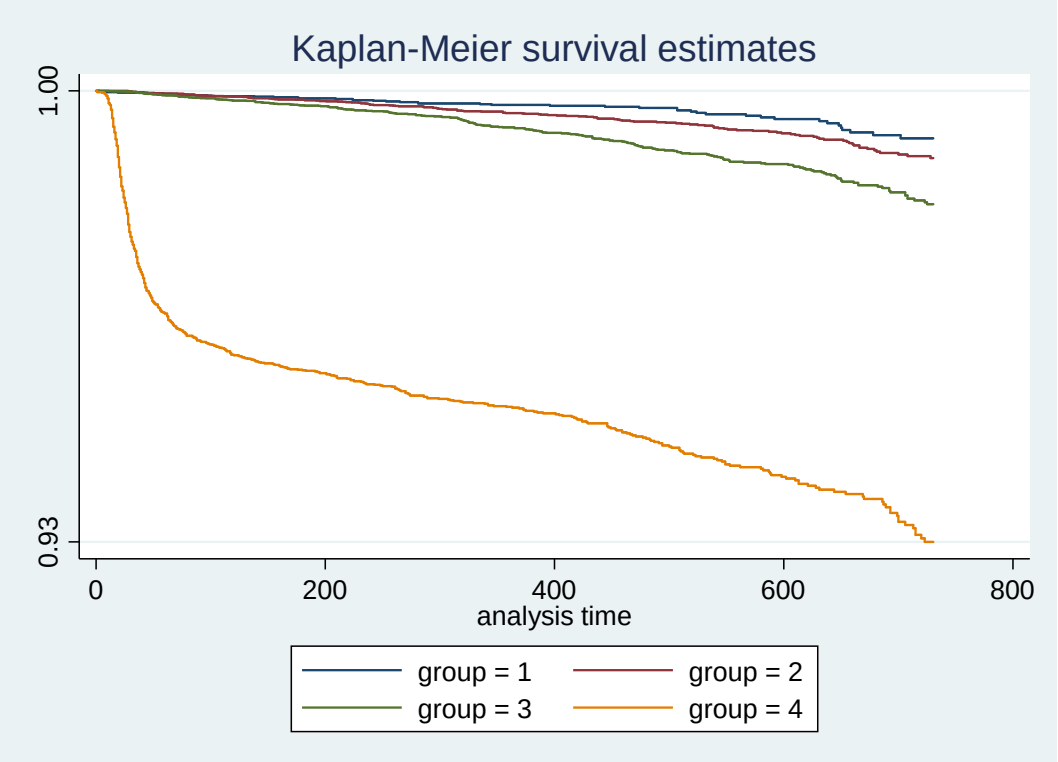


Figure 19: Kaplan Meier curves for 4 risk groups (in the screening cohort with positive and negative FOBTs), using the linear predictor which is divided into 4 using Cox's method – see methods section.

3.5.2 Adjusting for Optimism

The optimism of the model was assessed by calculating Van Houwelingen's heuristic shrinkage which was 0.995 ((3130.160 - 16)/ 3130.156). The linear predictor was then reassessed after applying this shrinkage factor and compared to the original linear predictor; this had a mean of 0.135 (SD: 0.685) and range -1.365 to 5.728. The calibration slope after applying the shrunken linear predictor was 1.005 (whereas it would have been 1.000 for the model which was not adjusted for optimism). **Figure 20** shows visually how the survival is adjusted after applying shrinkage for a high risk individual with a linear predictor of value 4.885 and a low risk individual with a linear predictor of -1.119.

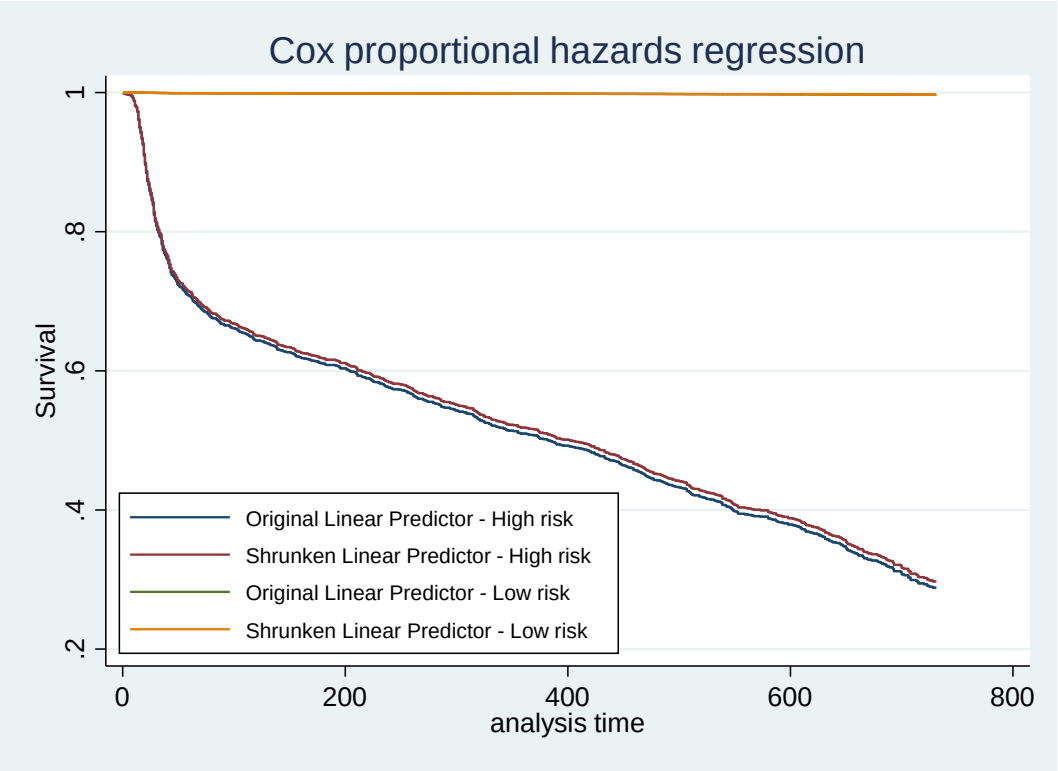


Figure 20: Survival for a high risk individual with a linear predictor of 4.885 which is shrunk to 4.860 and for a low risk individual -1.119 which is shrunk to -1.113. These individuals are from the screening cohort with positive and negative FOBTs.

To adjust performance statistics for optimism, internal validation was performed using 100 bootstrap replications for the C statistic, c-slope, D statistic and R^2 to quantify optimism. The optimism adjusted values for these performance parameters are displayed in **Table 14**. The bootstrapped uniform shrinkage factor (based on the optimism adjusted c-slope value) is slightly less (0.991) compared to the heuristic shrinkage (0.995). There was minimal optimism adjustment most likely due to the large sample size.

Statistic	Apparent Performance	Optimism (100 bootstrap replications)	Optimism adjusted performance (apparent minus optimism)
C statistic	0.854	0.004	0.850
c-slope	1.000	0.009	0.991
D statistic	2.344	0.046	2.298
R^2	0.568	0.010	0.558

Table 14: Optimism calculated for the C statistic, c-slope, D statistic and R^2 for the multivariable model developed using the screening cohort with positive and negative FOBTs. This uses 100 bootstrap replications and presents the corresponding optimism adjusted performance values. For bootstrap replications, the seed was set as '231398' in Stata.

3.5.3 Predicted Probabilities

This analysis was performed to determine individual risk probabilities from the model and to determine the distribution of risk in the sample population based on the predictors in the multivariable model.

The baseline survival for the Cox model was estimated non-parametrically at 2 years using the zero covariate value using the methods implemented in Stata. This was used along with the shrunken linear predictor to obtain the predicted probability of an individual being diagnosed with colorectal cancer/polyps at 2 years.

Before shrinkage, the baseline survival at 2 years was 0.9906 after shrinkage the baseline survival was 0.9905. The change in baseline survival over the 2 year period is shown below for the original and shrunken baseline survival (**Figure 21**). The shrunken baseline hazard is estimated using the heuristic linear predictor.

To generate risk probabilities both the heuristic linear predictor and the corresponding shrunken baseline survival were used for the final risk equation. The mean probability of being diagnosed with CRC or polyp within 2 years was 0.019 with a standard deviation of 0.058 (Range: 0.002, 0.946). The final risk equation for a high risk participant is shown in **Equation 2**.

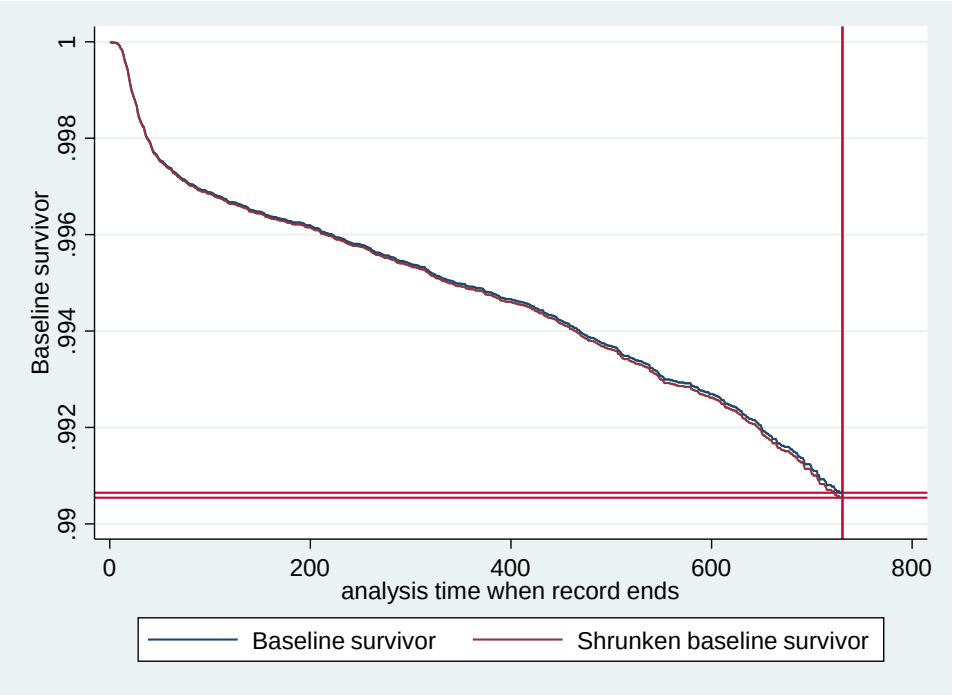


Figure 21: Baseline survivor versus the shrunken baseline survivor. The shrunken baseline survival at 2 years was estimated by setting the shrunken linear predictor as an offset and predicting the subsequent baseline survival. These results were derived from the screening cohort with positive and negative FOBTs.

Survival Probability

$S(2) = S_0(2)^{\exp(LP)}$

Where LP is the linear predictor and $S_0(2)$ is the baseline survival at 2 years.

Event Probability

P = 1 – S(2)

High risk Participant Example:

Survival Probability:

$0.29335536 = 0.9905431^{\exp(4.8603225)}$

Event Probability:

0.70664464 = 1-0.29335536

The probability of being diagnosed with CRC in a 2-year period for a high risk individual is 0.71.

Full Equation:

Survival Probability

$$S(2) = 0.991^{\exp(3.74x_1 + 0.21x_2 + 0.32x_3 + (-0.72)x_4 + 0.65x_5 + 0.85x_6 + 0.34x_7 + 0.08\left(\ln\left(\frac{x_8+1}{100}\right)+2.25\right) + 0.77x_9 + 0.20x_{10} + 0.27x_{11} + (-0.20)x_{12} + 0.03(x_{13}-66.97) + 0.91x_{14} + 0.09(x_{13}-66.97)x_7 + (-0.04)(x_{13}-66.97)x_1)}$$

0.991 = the baseline survival at 2 years $S_0(2)$
Where S(2) is the survival probability at 2 years (probability of not being diagnosed with colorectal cancer/polyps)

Event Probability
P = 1 – S(2)

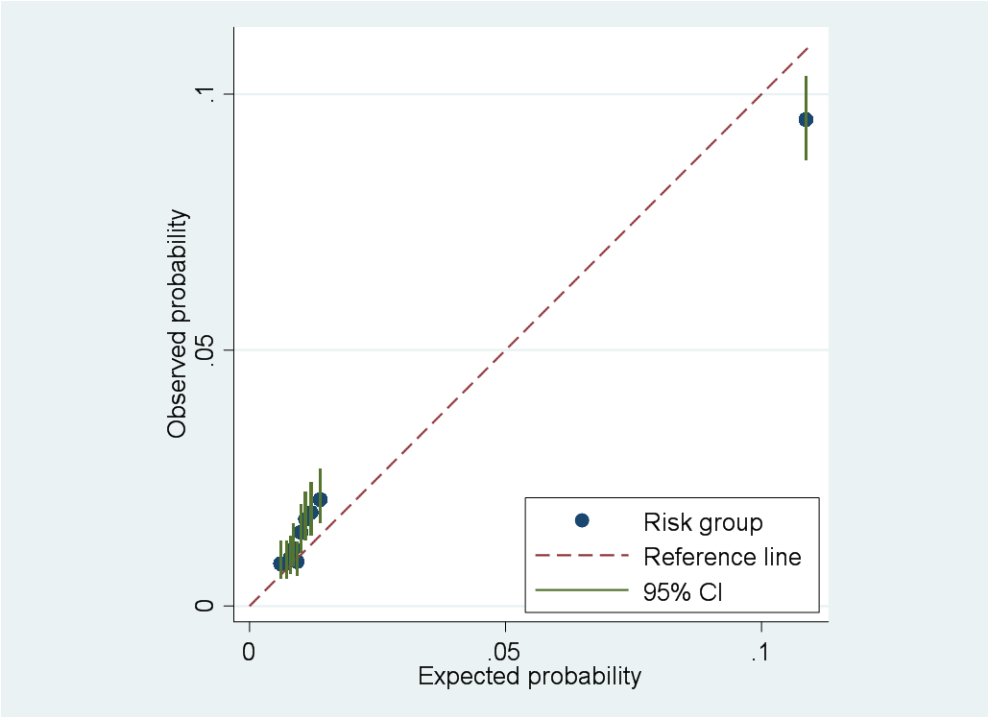
Where P is the probability of colorectal cancer/polyp being diagnosed within 2 years of the latest FOBT date; x_1 Latest FOBT result; x_2 ex-smoker; x_3 current smoker; x_4 Crohn's disease; x_5 previous polyps; x_6 flatulence; x_7 MCV <80fL; x_8 alcohol consumption; x_9 family history of gastrointestinal cancer; x_{10} abdominal pain/antispasmodic prescription; x_{11} diarrhoea; x_{12} sex; x_{13} age at FOBT, x_{14} Change in bowel habit.

Equation 2: Final risk equation for the screening cohort with positive and negative FOBTs with the shrunken baseline survival and using shrunken linear predictor values to correct for optimism.

3.5.4 Calibration

A calibration curve for the multivariable model adjusted for optimism is presented below for deciles of risk (**Figure 22**). For individuals at lower risk the model slightly underestimates the level of risk, whilst for the top risk group the model slightly overestimates the level of risk.

The separation between the risk groups gives an indication of how well the model discriminates between those with the disease and those without. The first 9 groups are spaced closely together with the mean probability of the 10th group being far removed. This is most likely due to whether an individual has either a positive or negative FOBT. Those with a positive FOBT are designated at much higher risk.



Risk Group	Calibration Expected Probability	Calibration Observed Probability	Calibration observed lower bound	Calibration observed upper bound
1	0.006	0.008	0.013	0.005
2	0.007	0.008	0.013	0.005
3	0.008	0.009	0.014	0.006
4	0.009	0.011	0.016	0.008
5	0.009	0.009	0.013	0.006
6	0.010	0.014	0.020	0.011
7	0.011	0.017	0.022	0.013
8	0.012	0.018	0.024	0.014
9	0.014	0.021	0.027	0.016
10	0.109	0.095	0.104	0.087

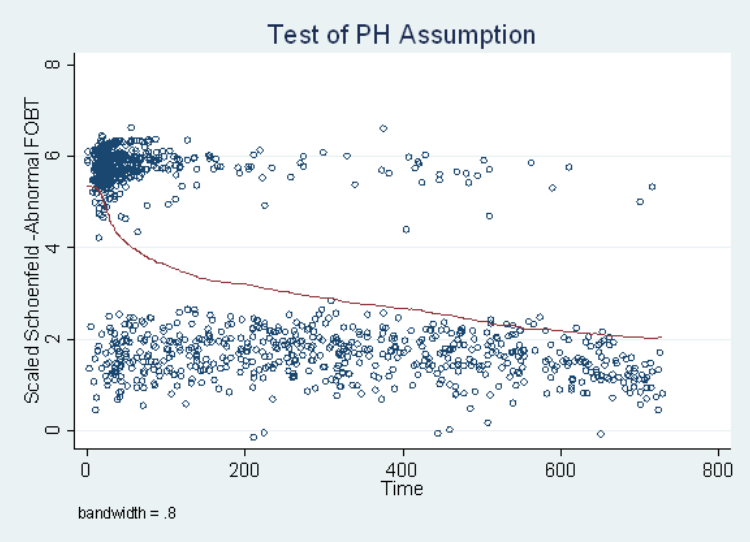
Figure 22: Calibration plot of observed probability versus expected probability using the multivariable model. The corresponding risk groups for each decile of probability are presented in the table below the figure.

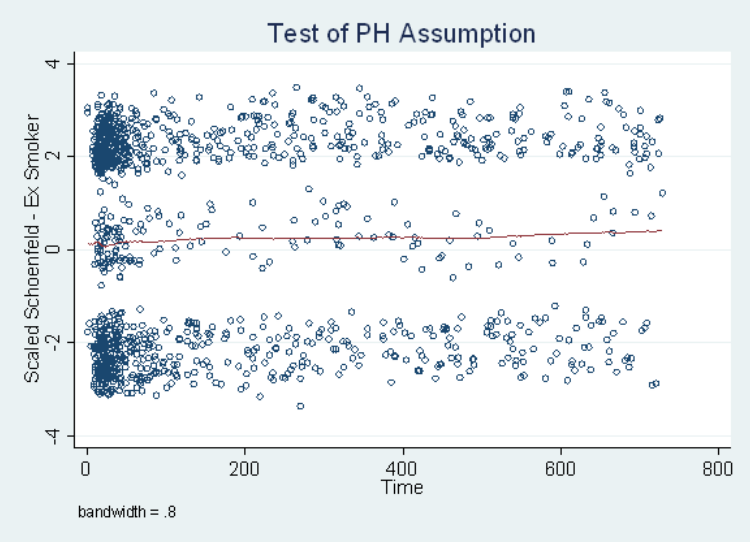
3.5.5 Cox Regression Diagnostics

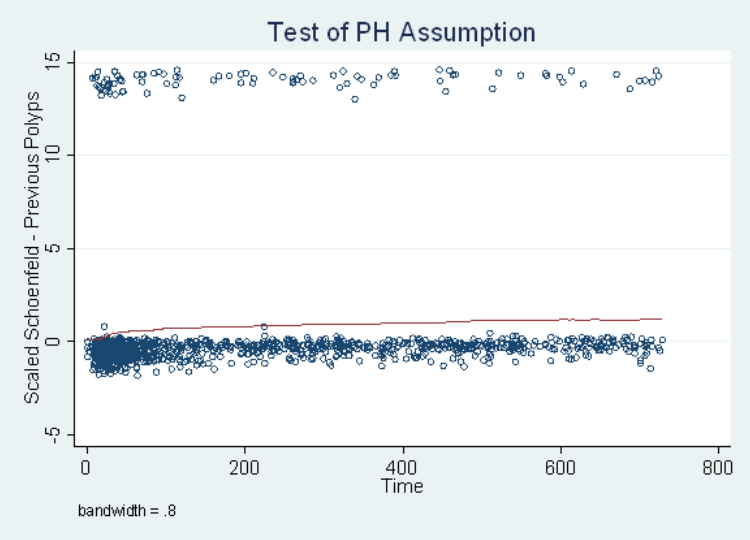
To test the proportional hazards assumption of the model, Schoenfeld residuals of the covariates were examined (**Appendix 7**). Significant results which had a p-value of less than 0.05 (which suggest they potentially violate the proportional hazards assumption) included; FOBT result (positive), smoking status current and ex-smoker, Crohn's disease, previous polyps and age at FOBT. The global test for the model also had a p-value <0.0001. Since this research uses a reasonably large dataset there is a lot of power to detect small deviations therefore graphical methods of proportional hazards were also assessed.

The scaled Schoenfeld residuals were plotted for these covariates where a straight horizontal line supports that there is not a violation of the proportional hazards assumption (**Figure 23**). The lines are reasonably straight but do deviate slightly over time for example with previous polyps there is a slight increase over time and with ex-smokers. The abnormal FOBT result due to an increased number of earlier events has a steep decline and then levels out over time.

Log-log plots were also plotted for these variables to assess proportionality by determining whether the lines are roughly parallel (**Figure 24 & Figure 25**). The latest FOBT result is roughly parallel over the time period due to the cut off of a two year follow up. Crohn's disease does not appear to have proportional hazards and there are crossovers with previous polyps and smoking status.


Test of PH Assumption
Scaled Schoenfeld -Abnormal FOBT
0
2
4
6
8
0
200
400
600
800
Time
bandwidth = .8


Test of PH Assumption
Scaled Schoenfeld - Ex Smoker
-4
-2
0
2
4
0
200
400
600
800
Time
bandwidth = .8


Test of PH Assumption
Scaled Schoenfeld - Previous Polyps
-5
0
5
10
15
0
200
400
600
800
Time
bandwidth = .8

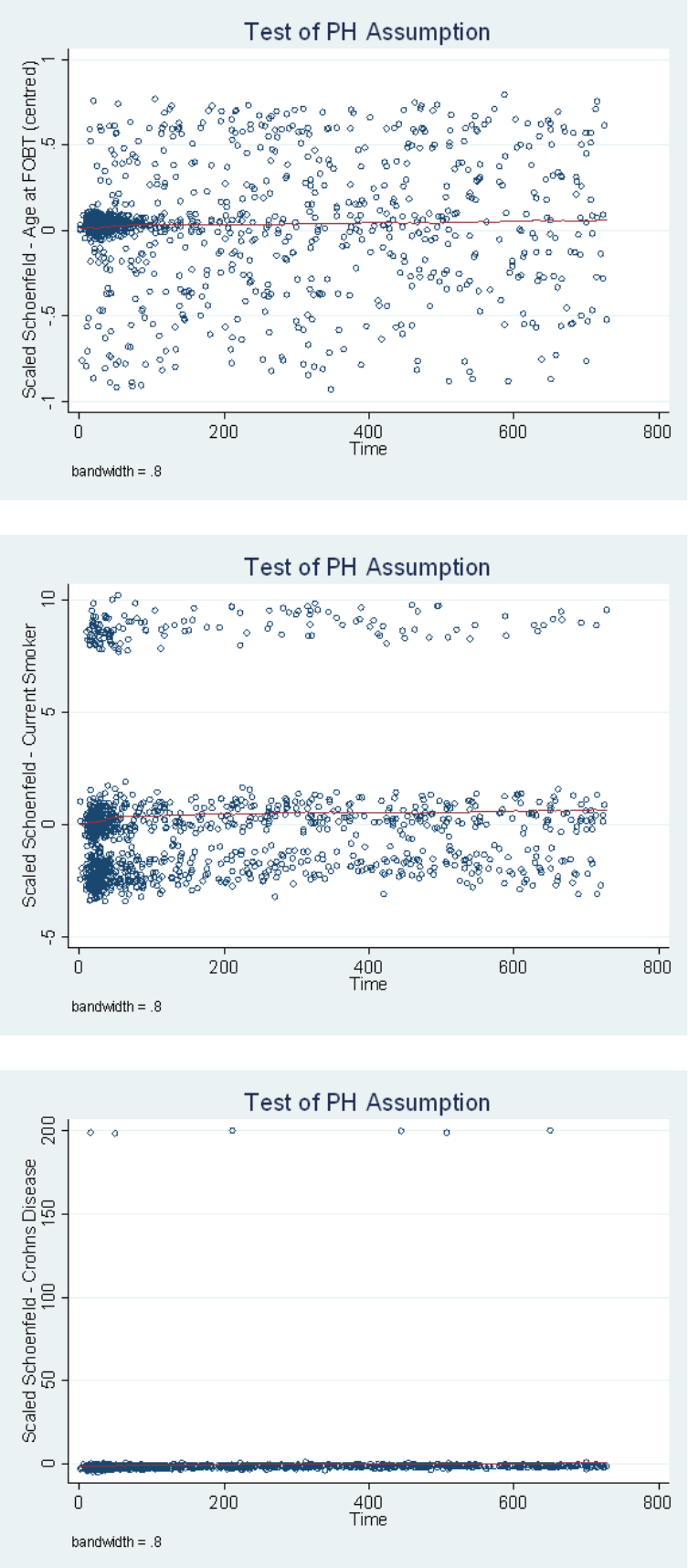


Figure 23: Schoenfeld residual plots for variables which had a p value of <0.05 when testing the proportional hazards assumption in the derived screening cohort. These variables included: Positive FOBT (1st plot), ex-smoker (2nd plot), previous polyps (3rd plot), age at FOBT (4th plot), current smoker (5th plot), Crohn's disease (6th plot).

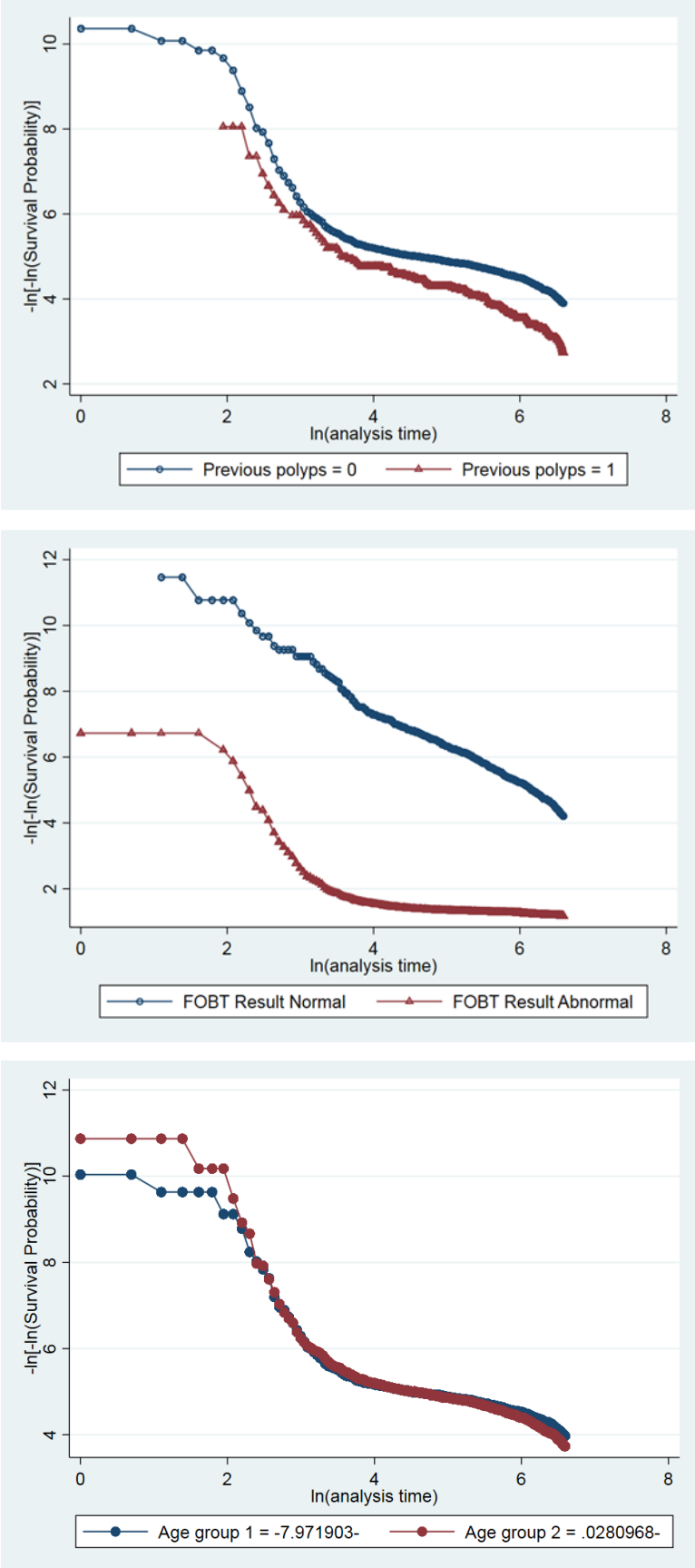


Figure 24: Log-log plots to test the Cox proportionality assumption for previous polyps (1st plot), FOBT result (2nd plot), and Age group (3rd plot - which was split into 2 equally sized groups) for the derived screening cohort).

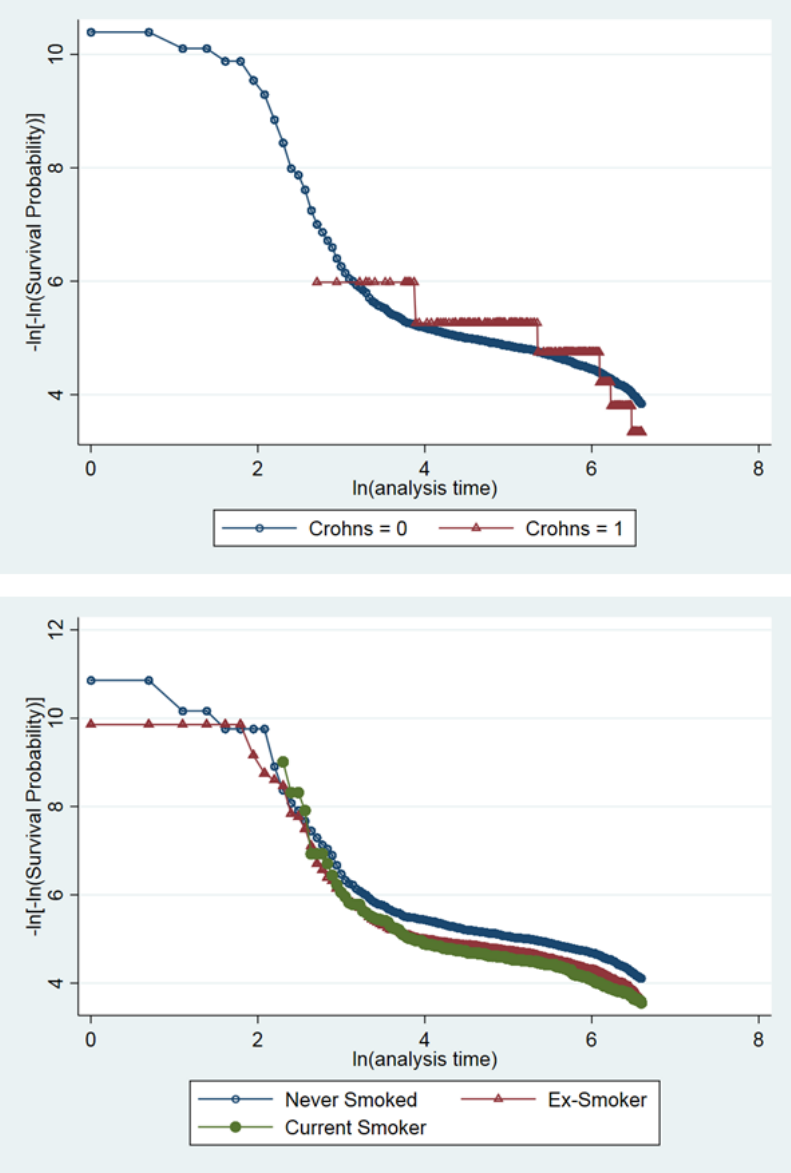


Figure 25: Log-log plot to test the Cox proportionality assumption for Crohn's disease (1st plot) and Smoking status (2nd plot) for the derived screening cohort.

Overall model fit was assessed using Cox-Snell residuals by plotting the Nelson-Aalen cumulative hazard function against Cox-Snell residuals (**Figure 26**).[79] This plot shows that the cumulative hazard function does not have an exponential distribution with hazard rate of one since the Cox-Snell residuals deviate from this line particularly at the tail end where there are fewer events. This suggests that perhaps a more flexible parametric model may have a better fit.

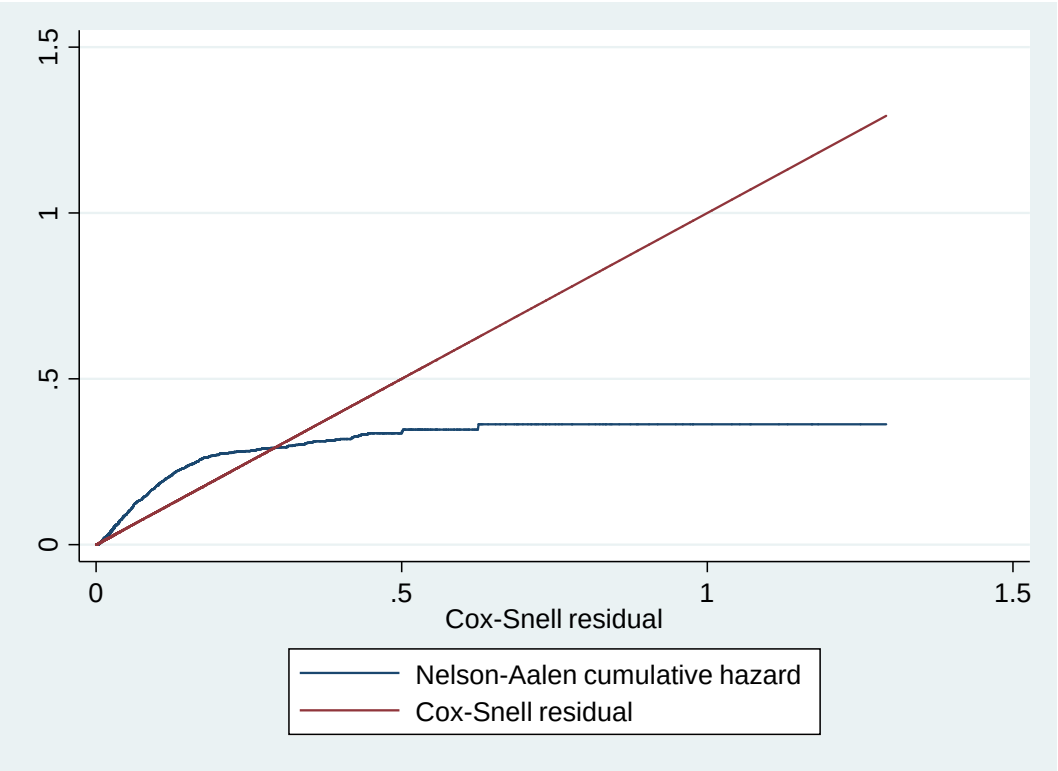


Figure 26: Assessment of overall model fit for the model derived from the screening cohort with positive and negative FOBTs using Cox-Snell residuals by plotting the Nelson-Aalen cumulative hazard function against Cox-Snell residuals. For a good model fit, the cumulative hazard function should follow the Cox-Snell residuals.

3.5.6 Parametric Survival Models

Parametric models were investigated as an extension to Cox Regression to determine whether these types of model gave a better fit to the data and therefore more accurate parameter estimates. Parametric models also offer more with post-estimation, allowing predictions to be made at multiple time points. The generalised gamma, loglogistic and log normal parametric models use the accelerated time metric and can be used to derive time ratios which are arguably more interpretable compared to hazard ratios.[34] The Weibull, exponential and Gompertz parametric models on the other hand have a proportional-hazards parameterisation. The AIC was used to compare the different parametric survival models since the models do not require to be nested as with likelihood ratio testing (**Table 15**). The smallest AIC and therefore the model which may fit the data best is the generalised gamma model. Generalised gamma models are useful when the hazard rises to a peak before dropping.[40] This is a reasonable assumption in the setting here and could be investigated further. Furthermore, the generalised gamma distribution is a three-parameter distribution which has a flexible hazard function allowing for many types of shape.[60] The fit was investigated further by plotting Cox-Snell residuals (**Figure 27**), Nelson

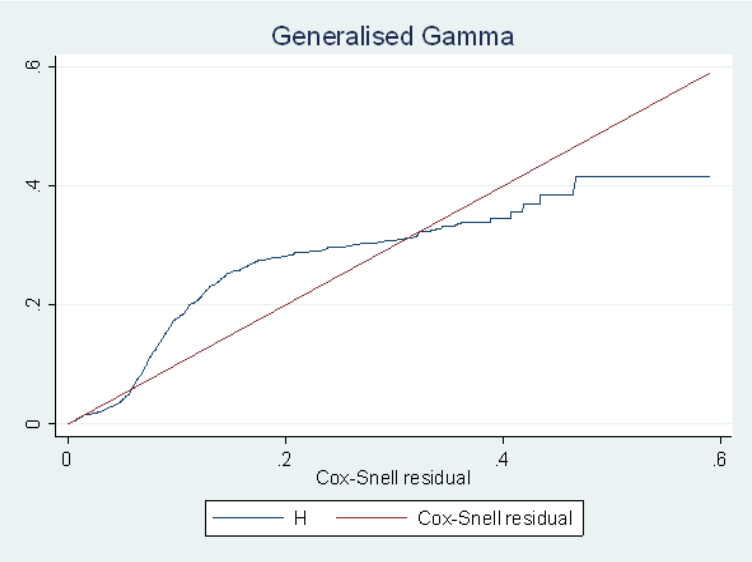
Aalen cumulative hazard plots and Kaplan Meier function graphs for the parametric models to assess the fit visually.

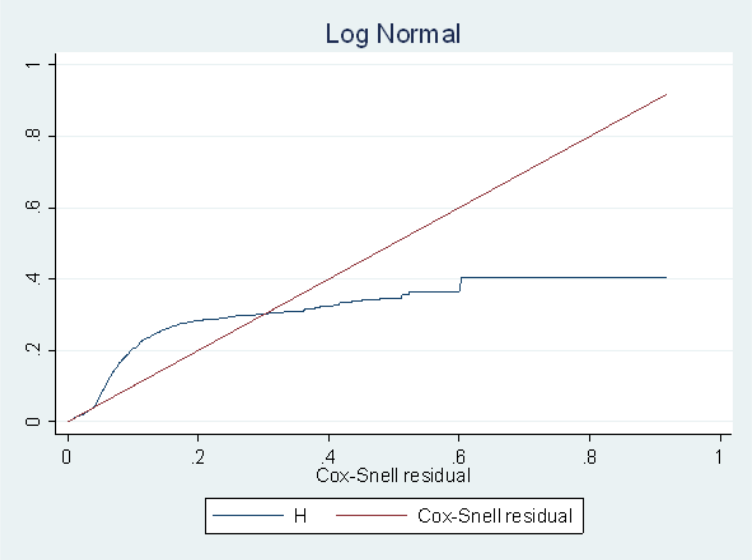
Model	Observations	Log likelihood null	Log likelihood model	Degrees of freedom	AIC	BIC
Exponential	1,197	-8233.44	-6649.63	17	13333.25	13419.74
Weibull	1,197	-8071.73	-6490.23	18	13016.47	13108.04
Gompertz	1,197	-8141.97	-6545.09	18	13126.18	13217.75
Lognormal	1,197	-8048.78	-6241.59	18	12519.19	12610.76
Loglogistic	1,197	-8071.51	-6427.32	18	12890.65	12982.22
Generalised gamma	1,197	-8071.95	-6159.85	19	12357.7	12454.36

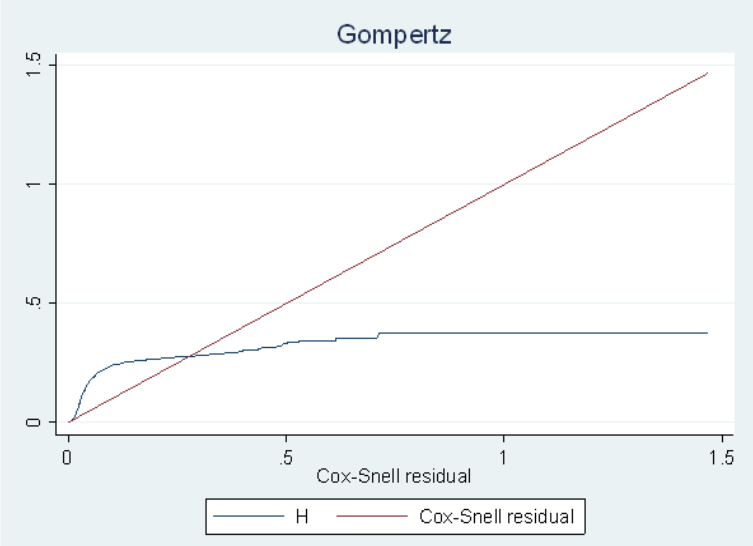
Table 15: Difference between the different parametric models for the model derived from the screening cohort with positive and negative FOBTs compared with the semi-parametric Cox regression model.

Although all models show deviation from the reference line, the generalised gamma model has a better fit at the tail end of the data. The Nelson Aalen cumulative hazard plots were assessed for all the parametric models, there was similar fit to the data for all the models but the generalised gamma model had slightly better fit overall **Figure 28**. The Kaplan Meier function graphs show very similar fit for Weibull, generalised gamma, lognormal and loglogistic models (**Figure 29)**.

The coefficients for the generalised gamma model are reported in **Table 16**.


Generalised Gamma
0
.2
.4
.6
Cox-Snell residual
H
Cox-Snell residual


Log Normal
0
.2
.4
.6
.8
1
Cox-Snell residual
H
Cox-Snell residual


Gompertz
0
.5
1
1.5
Cox-Snell residual
H
Cox-Snell residual

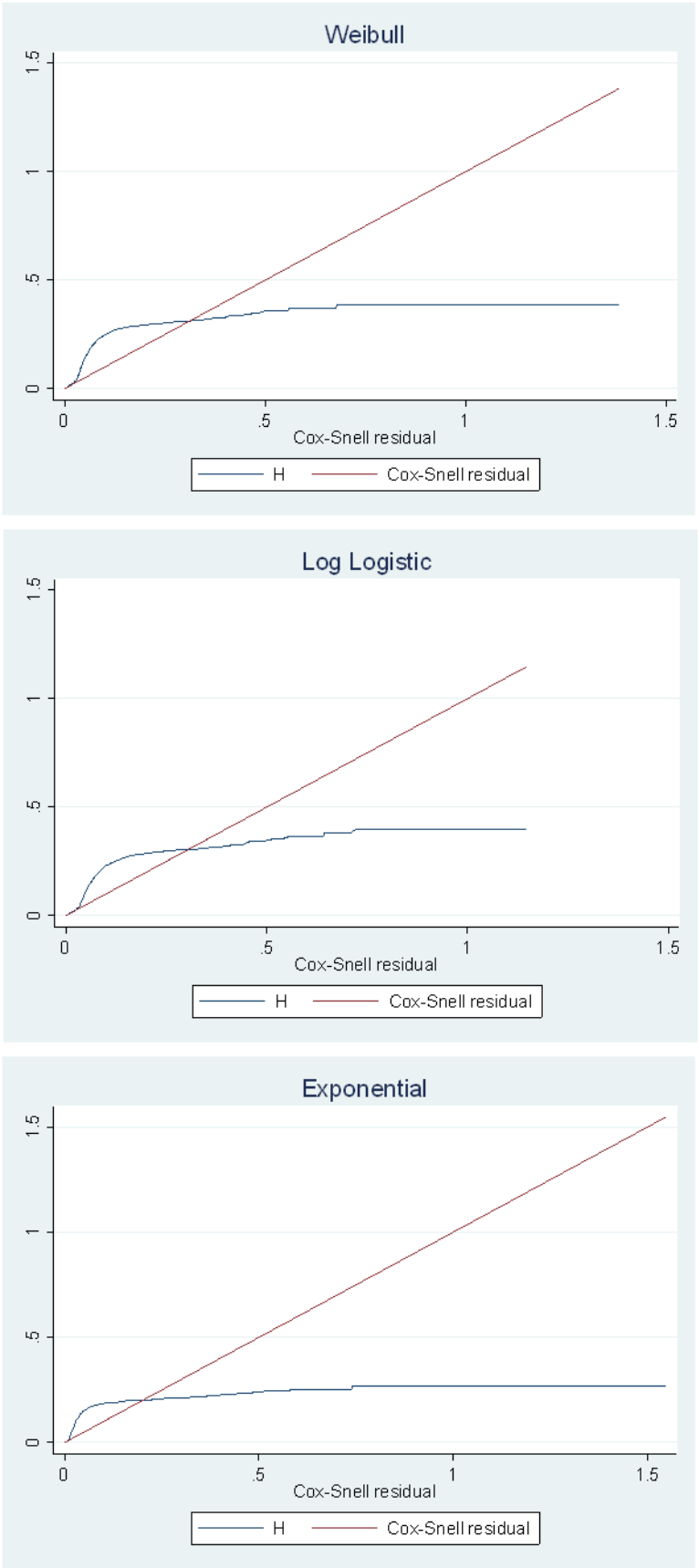
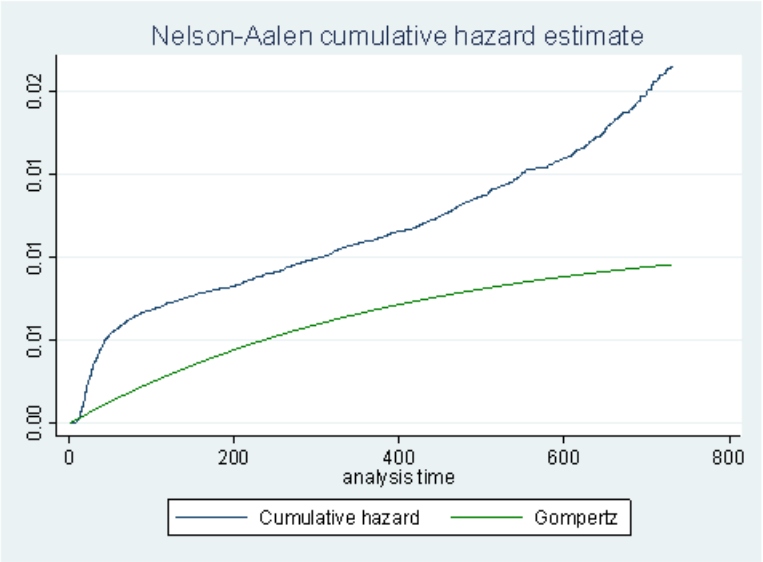
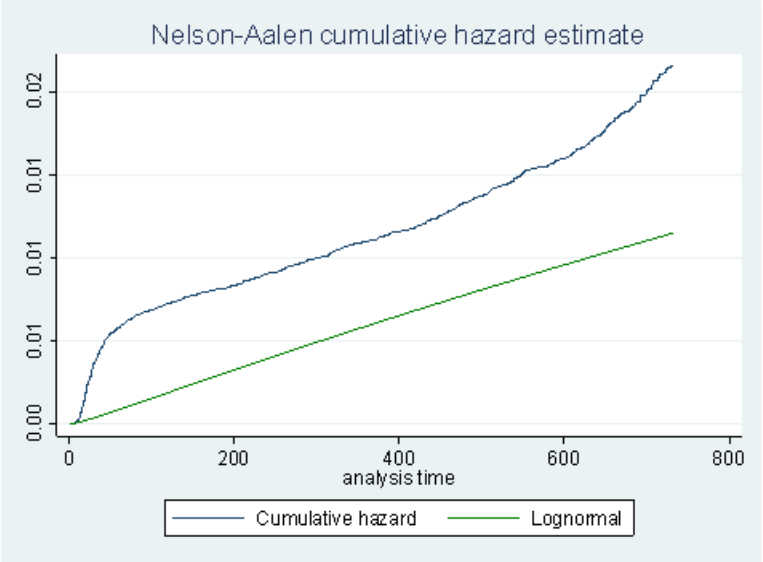
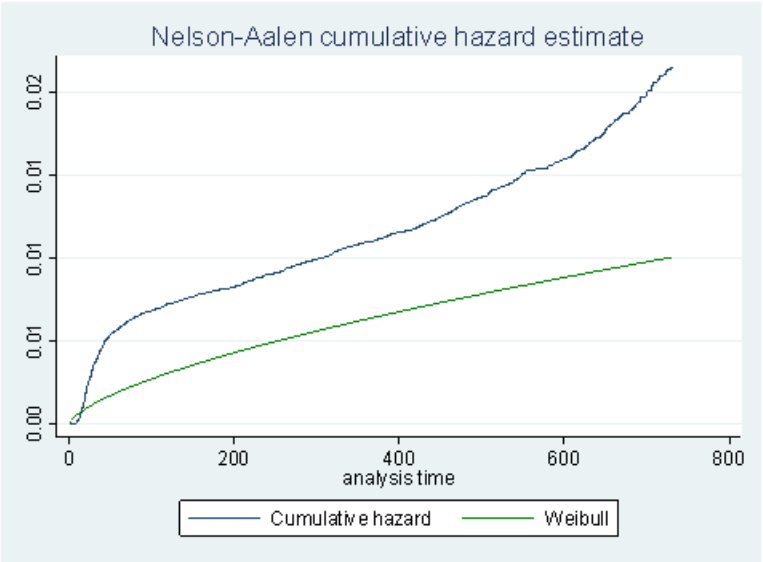


Figure 27: Cox Snell Residuals plotted for all considered parametric models, for the multivariable model derived from the screening cohort with positive and negative FOBTs, to assess model fit.


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0.02
0
200
400
600
800
analysis time
Cumulative hazard
Gompertz


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0.02
0
200
400
600
800
analysis time
Cumulative hazard
Lognormal


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0.02
0
200
400
600
800
analysis time
Cumulative hazard
Weibull

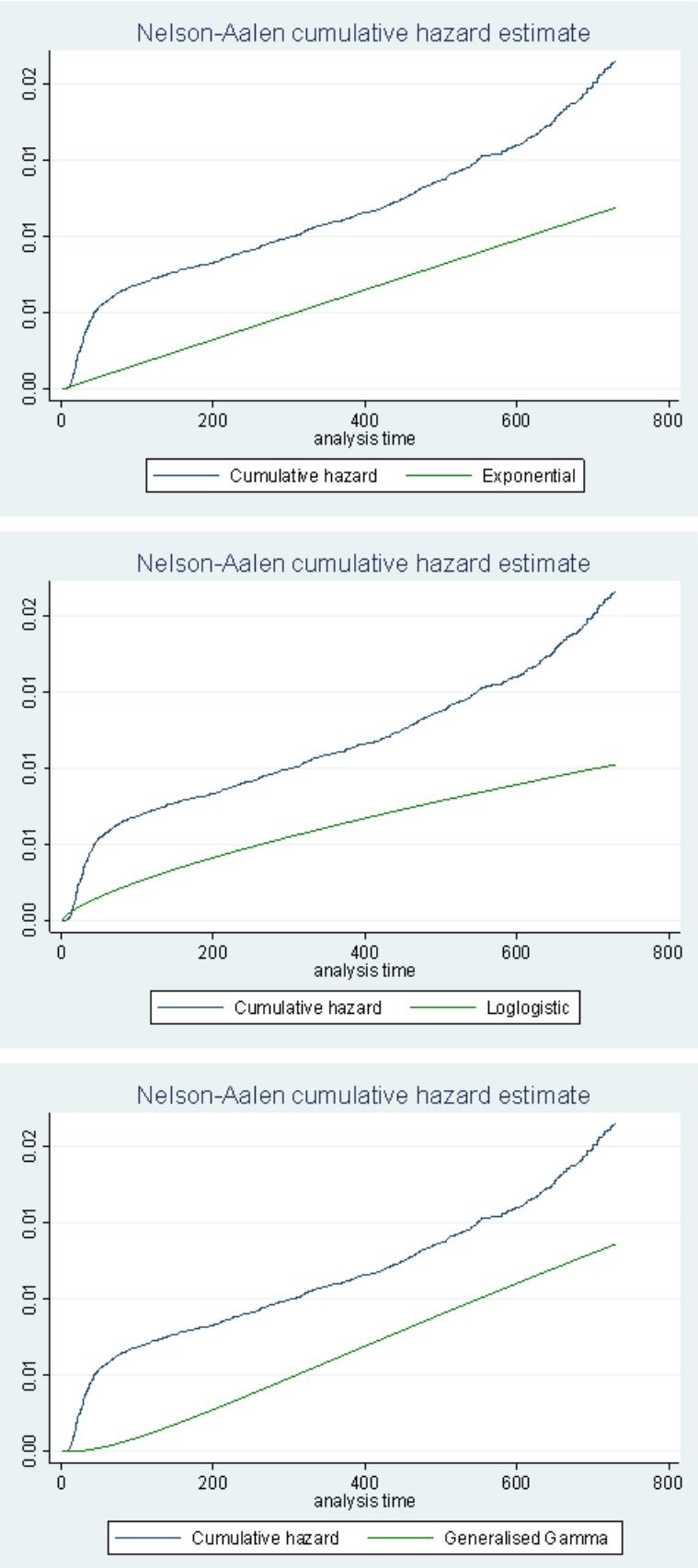
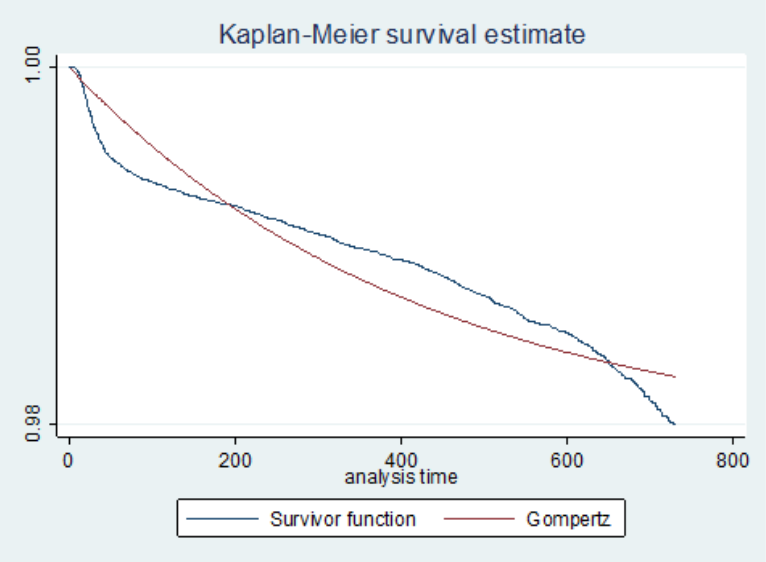
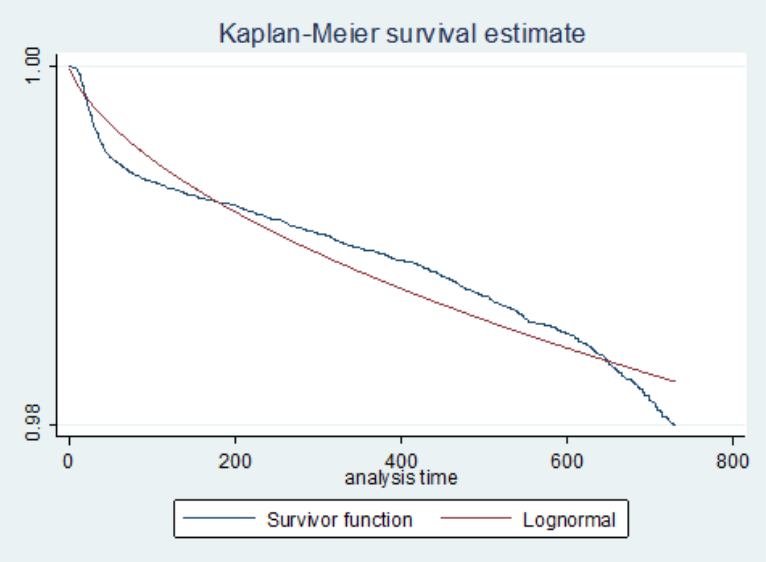
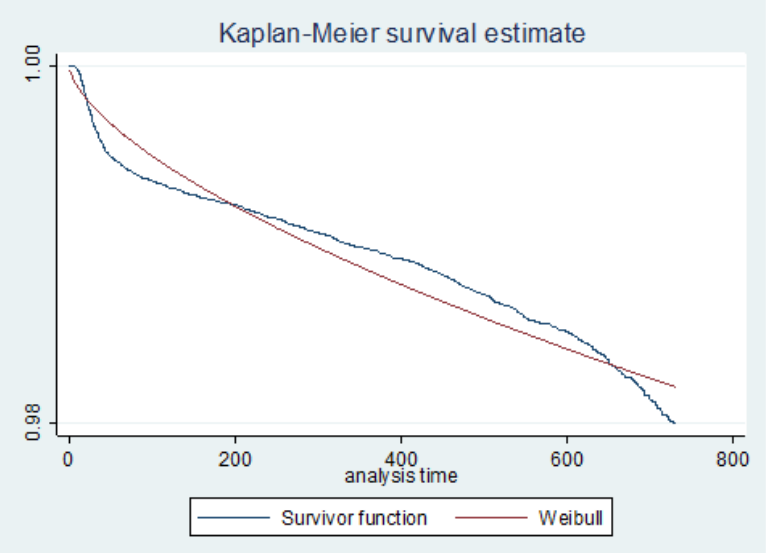


Figure 28: Nelson-Aalen cumulative hazard plots for all considered parametric models to assess model fit of the multivariable model derived from the screening cohort with positive and negative FOBTs.


Kaplan-Meier survival estimate
1.00
0.98
0
200
400
600
800
analysis time
Survivor function
Gompertz


Kaplan-Meier survival estimate
1.00
0.98
0
200
400
600
800
analysis time
Survivor function
Lognormal


Kaplan-Meier survival estimate
1.00
0.98
0
200
400
600
800
analysis time
Survivor function
Weibull

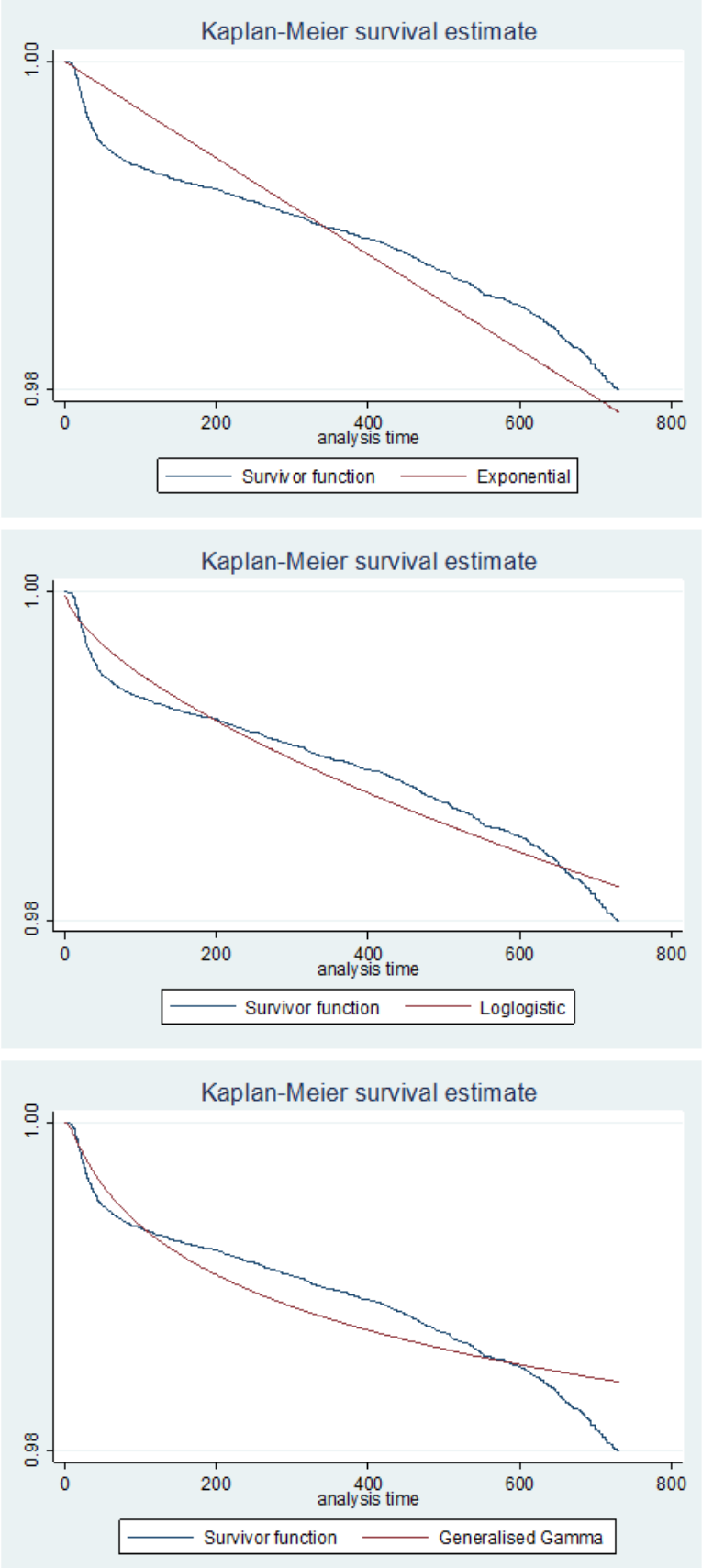


Figure 29: Kaplan-Meier function graphs for all considered parametric models to assess model fit of the multivariable model derived from the screening cohort with positive and negative FOBTs.

3.5.7 Model Performance measures for the best fitting parametric models

Based on the plots in the previous section and the AIC value, the generalised gamma model had the best fit to these data. The appropriateness of the lognormal and Weibull models were also assessed by determining whether the second ancillary parameter (kappa) was equal to 0 or 1.[60 75] This was significant for kappa equal to 1 suggesting the Weibull model may also have an appropriate fit ($p<0.001$). Harrell's C-statistic was very similar across all three parametric models; generalised gamma 0.859 (95% CI: 0.845, 0.872), Weibull 0.854 (95% CI: 0.841, 0.868) and lognormal 0.857 (95% CI: 0.844, 0.871). These values were also comparable to the equivalent Cox Regression model (0.854 (95% CI: 0.841, 0.868)). The coefficients for the models are presented below in **Table 16** to aid comparison, along with the discrimination and other performance measures where applicable. The Weibull model is presented both in the accelerated failure time and proportional hazards parameterisations for comparison between the other accelerated failure time models and Cox regression (which uses the proportional hazards assumption).

Variable	Weibull (AFT) coefficients	Weibull (PH) coefficients	Lognormal (AFT) coefficients	Generalized gamma (AFT) coefficients	Cox (PH) coefficients
MCV*age at FOBT interaction	-0.132	0.086	-0.130	-0.105	0.086
FOBT result*age at FOBT interaction	0.059	-0.039	0.045	0.041	-0.037
FOBT Result (positive)	-5.766	3.765	-5.595	-5.656	3.741
Smoking Status:					
ex-smoker	-0.313	0.204	-0.276	-0.213	0.206
current smoker	-0.496	0.324	-0.462	-0.411	0.323
Crohn's Disease Diagnosis Recorded	1.126	-0.735	0.933	0.562	-0.722
Previous Polyps Diagnosed	-0.986	0.644	-1.075	-1.043	0.648
Flatulence Symptom Recorded	-1.270	0.829	-1.064	-0.780	0.850
MCV <80fL	-0.514	0.336	-0.604	-0.653	0.344
Alcohol consumption (units per week)	-0.125	0.082	-0.097	-0.067	0.082
Family History of Gastrointestinal Cancer	-1.192	0.779	-0.925	-0.712	0.766
Abdominal pain/antispasmodic prescription recorded	-0.313	0.204	-0.351	-0.388	0.199
Diarrhoea symptom	-0.429	0.280	-0.531	-0.579	0.272
Sex	0.307	-0.200	0.249	0.191	-0.196
Age at FOBT	-0.049	0.032	-0.030	-0.019	0.033
Change in bowel habit symptom	-1.399	0.913	-1.095	-1.011	0.908
Constant	**13.839**	**-9.036**	**13.455**	**12.919**	-
Ancillary parameter	**0.653**	**0.653**	**2.911**	**4.459**	-
Kappa (Ancillary parameter 2 for the gamma model)	**NA**	**NA**	**NA**	**-1.209**	-
Log likelihood	**-6490.233**	-6490.233	-6241.594	-6159.848	-11733.936[a]
AIC	13016.470	13016.470	12519.190	12357.700	23499.87[a]
BIC	13108.040	13108.040	12610.760	12454.360	23581.27[a]
Harrell's C Statistic (95%CI)	0.854 (0.841,0.868)	0.854 (0.841,0.868)	0.857 (0.844, 0.871)	0.859 (0.845, 0.872)	0.854 (0.840, 0.868)
R^2	0.759	0.571	0.270	-[b]	0.568
D Statistic	3.628	2.361	0.971	-[b]	2.344
Adjusted R^2 (Bootstrap CI 100 replications)	0.755585 (0.721523 0.781859)	0.566074 (0.540843 0.594144)	0.260 (0.241,0.293)	-[b]	0.563 (0.526, 0.593)
Optimism adjusted Calibration Slope (also shrinkage factor for linear predictor)	**0.990**	0.991	0.996	-[c]	0.991
Optimism Harrell's C Statistic	**0.851**	0.851	0.853	-[c]	0.850
Optimism adjusted D statistic	**3.550**	2.313	0.949	-[b]	2.298
Optimism adjusted R2	**0.751**	0.561	0.261	-[b]	0.558

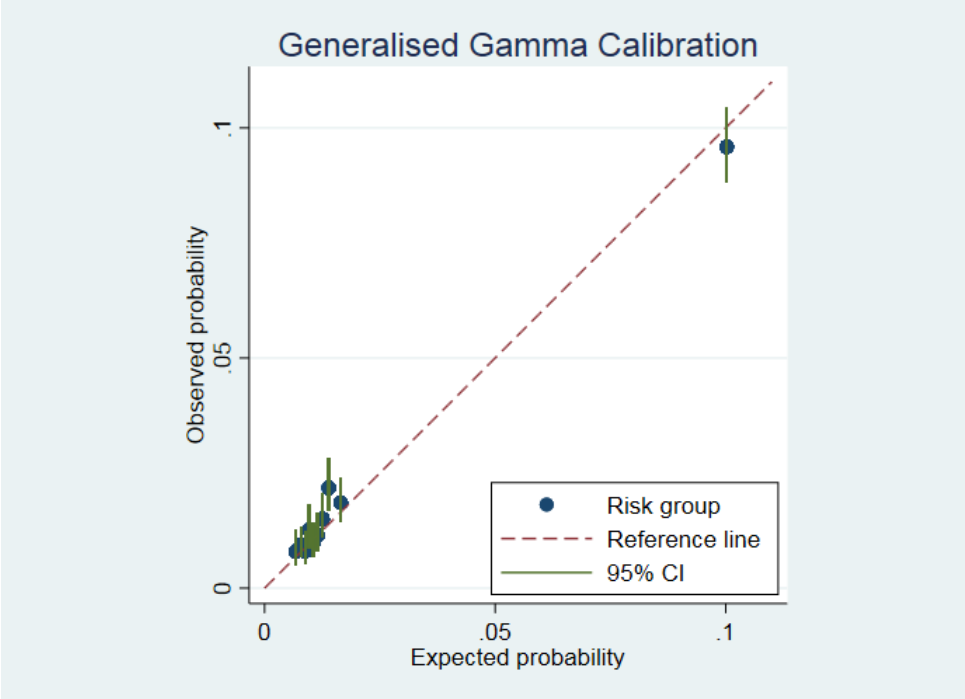
[a] The log likelihood/AIC/BIC from the Cox model is not comparable to parametric models since it uses partial likelihood whereas the other models use full maximum likelihood

[b] The R^2_D measure is not available for the generalised gamma distribution

[c] Optimism adjustment for the gamma model requires extensive computational time for bootstrapping in this dataset

Table 16: Comparison of the best fitting parametric models compared to the Cox model for a sample population with both negative and positive FOBT results. Model coefficients, model constants, ancillary parameters, AIC, BIC, R^2, D statistic and optimism adjusted performance metrics are presented for comparison. The R^2 used in this instance is Royston and Sauerbrei's (2004) R^2_D measure of explained variation for survival models based on their index of discrimination (D).[73] The adjusted R^2 measure also considers the number of covariates in the model. For non-proportional hazards models R^2 for explained variation is not interpretable but can be used as an index of determination.[74]

The calibration of the models as assessed through calibration plots (**Figure 31**) was also similar to the equivalent Cox Regression (**Figure 22**). The generalised gamma and the lognormal models had a similar pattern for the observed versus predicted risk groups and the Weibull model had a similar pattern as compared to the Cox Regression model. The generalised gamma model had the closest alignment to the 45 degree line (representing good calibration). Calibration would ideally be tested in an external validation scenario to determine whether significant recalibration was required in a different sample population.



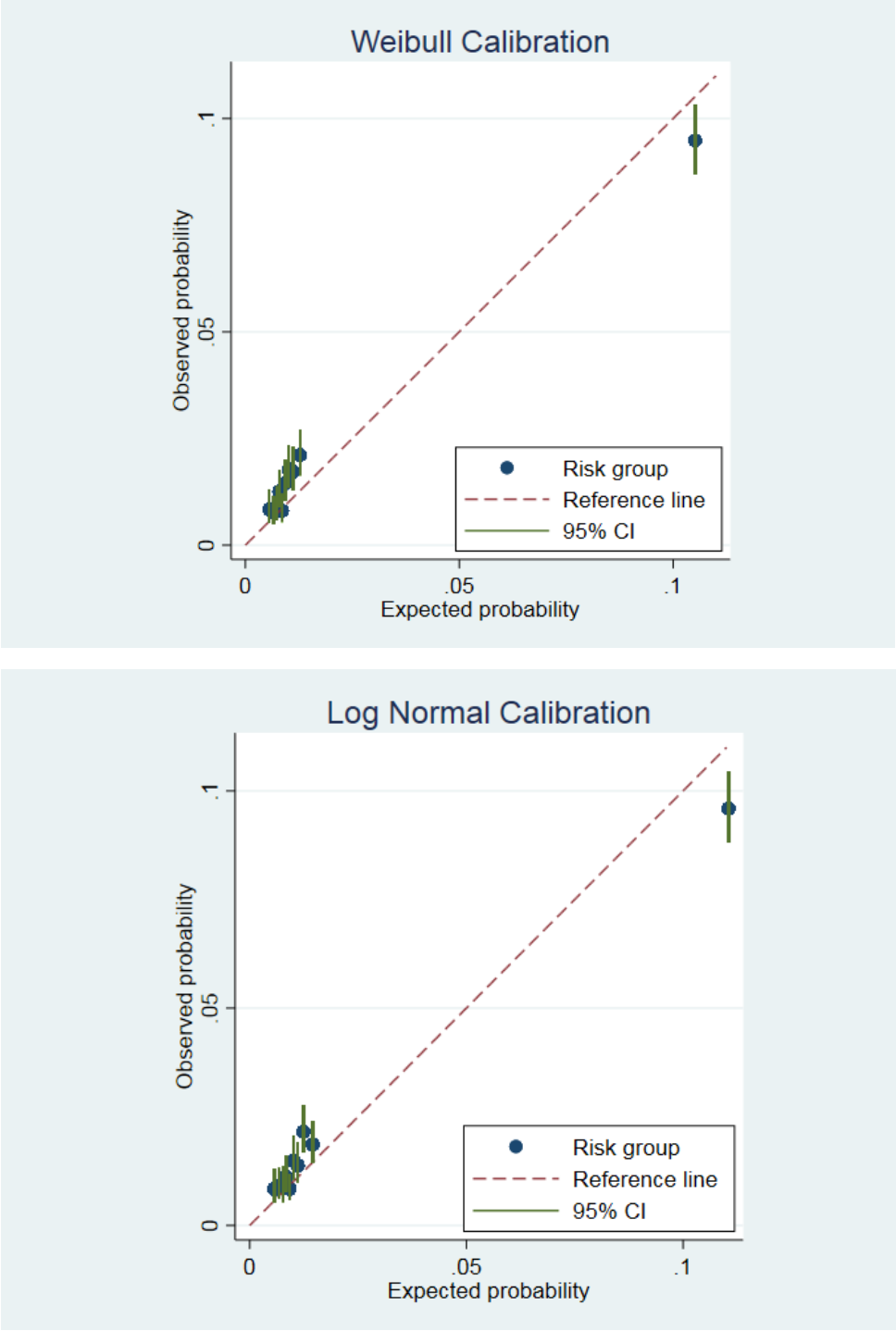


Figure 30: Calibration plot of observed probability versus expected probability using the Generalised Gamma model (top), Weibull model (middle) and lognormal model (bottom) for a sample population with both negative and positive FOBT results.

3.6 Multivariable Analysis Risk Prediction Model Development (n=95,792)

3.6.1 Cox Regression for Negative Results Only (n = 95,792)

This analysis allowed the identification of additional predictors which could be used in a screening population with negative results for screening referral decisions.

The FOBT has only a sensitivity of around 50% and so other factors could be used to determine whether a patient should be referred if they have had a negative result. The probabilities derived for the model developed in those with a positive or negative FOBT reflects the underlying screening pathway whereby those with a positive result are referred on quickly for diagnostic testing and so have a higher probability of diagnosis. Therefore, Cox regression was used to investigate if the additional information from the electronic GP record could be used to make better screening referral decisions for those with negative FOBT results. After setting the data up for survival analysis, there were 95,792 observations with 587 events and 37,154,249.5 total analysis time at risk. The median follow up time in days was 384 (95% CI: 381 to 386). Restricted mean diagnosis-free survival was 726.5 days. The original model had 31 degrees of freedom with all predictors considered (39 when including interactions).

Model building used the same 'mfp' function as described in **Section 3.5** using a p value of 0.05 for backwards elimination, for testing between multivariable fractional polynomials for continuous variables and for interactions. The final model (**Table 17**) after assessing all eligible predictors included; smoking status, whether a patient had an IBS diagnosis, previous polyps diagnosed, flatulence, weight loss, MCV of <80fL compared to a MCV of ≥80fL, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit. There were no significant interactions (interactions investigated included; smoking status and age, smoking and sex, MCV and sex, MCV and age, age and sex) and a linear model had the best fit to continuous predictors during multivariable analysis (age at FOBT was centred).

Apparent model performance parameters included Harrell's C statistic which was 0.658 (95% CI: 0.633, 0.683) and Somers D 0.316. Harrell's C statistic means that the predictors used in the model correctly identify the order of survival times for pairs of patients 66% of the time. The final model had 13 degrees of freedom with an AIC of 12493.68 and BIC of 12550.56 (N=587 when calculating BIC)). Overall model fit was assessed using adjusted R^2

which was 0.151 (bootstrapped CI 100 reps: 0.122, 0.204.[73 78] Regular R^2 was 0.164 (95% CI: 0.126, 0.204) with D statistic of 0.906.

Variable	Observed Coefficient	Bootstrapped Standard Error	z	P>z	[95% Confidence Intervals]	
Smoking Status:						
ex-smoker	0.285	0.105	2.720	0.006	0.080	0.491
current smoker	0.516	0.150	3.450	0.001	0.223	0.810
IBS	0.258	0.125	2.070	0.039	0.014	0.502
Previous Polyps Diagnosed	1.225	0.132	9.250	0.000	0.965	1.484
Flatulence Symptom Recorded	0.953	0.505	1.890	0.059	-0.037	1.944
Weight loss	0.867	0.343	2.530	0.011	0.195	1.539
MCV <80fL	0.877	0.291	3.010	0.003	0.306	1.447
Family History of Gastrointestinal Cancer	0.603	0.248	2.430	0.015	0.117	1.089
Abdominal pain/antispasmodic prescription recorded	0.365	0.126	2.890	0.004	0.117	0.612
Diarrhoea symptom	0.572	0.157	3.640	0.000	0.264	0.880
Sex	-0.323	0.077	-4.180	0.000	-0.475	-0.172
Age at FOBT	0.034	0.010	3.480	0.000	0.015	0.053
Change in bowel habit symptom	0.793	0.273	2.900	0.004	0.257	1.328

Table 17: Cox regression model (coefficients) after 'mfp' selection for patients with a negative FOBT only. The continuous variable age at FOBT has been centred (age_at_FOBT-66.97).

The linear predictor from this model had a mean of 0.151 and a standard deviation of 0.451 (range: -0.593 to 4.398). The distribution of the linear predictor is shown in **Figure 31**. Discrimination was also assessed by analysing the separation between Kaplan-Meier curves for 4 risk groups (where the linear predictor is divided into 4 groups). Separation was greater for those in the higher risk groups (group 3 and 4), but was smaller for risk group 1 and risk group 2 (**Figure 32**).

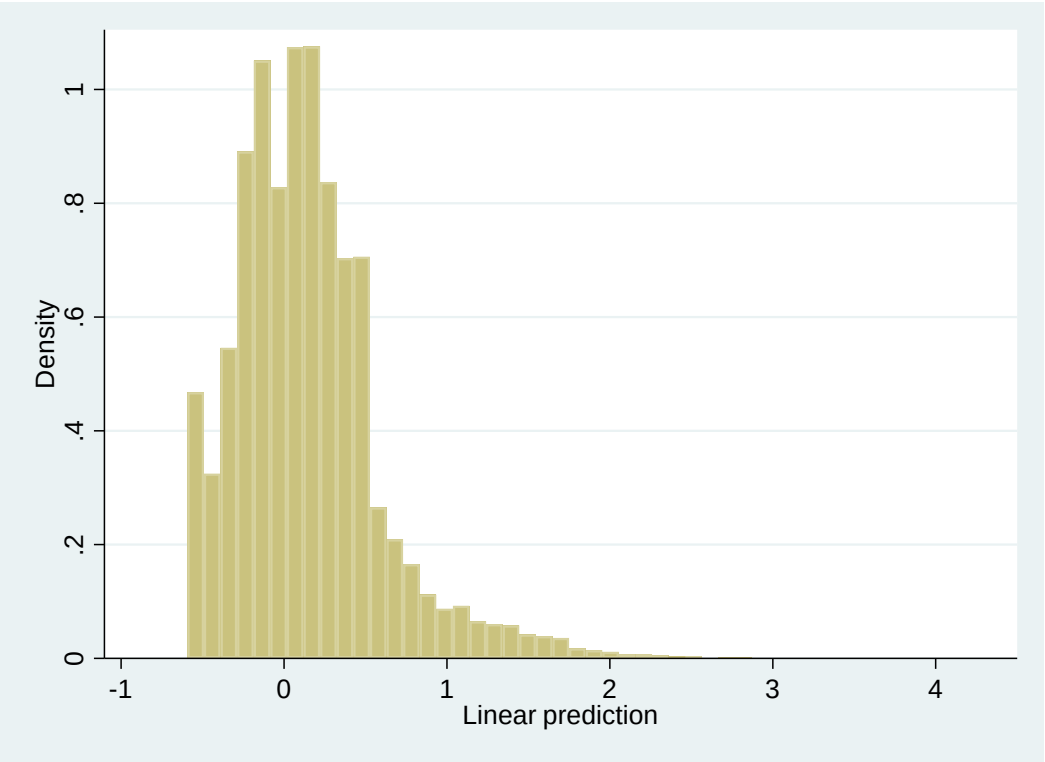


Figure 31: Distribution of the linear predictor for the final multivariable model derived from a population with negative FOBTs only.

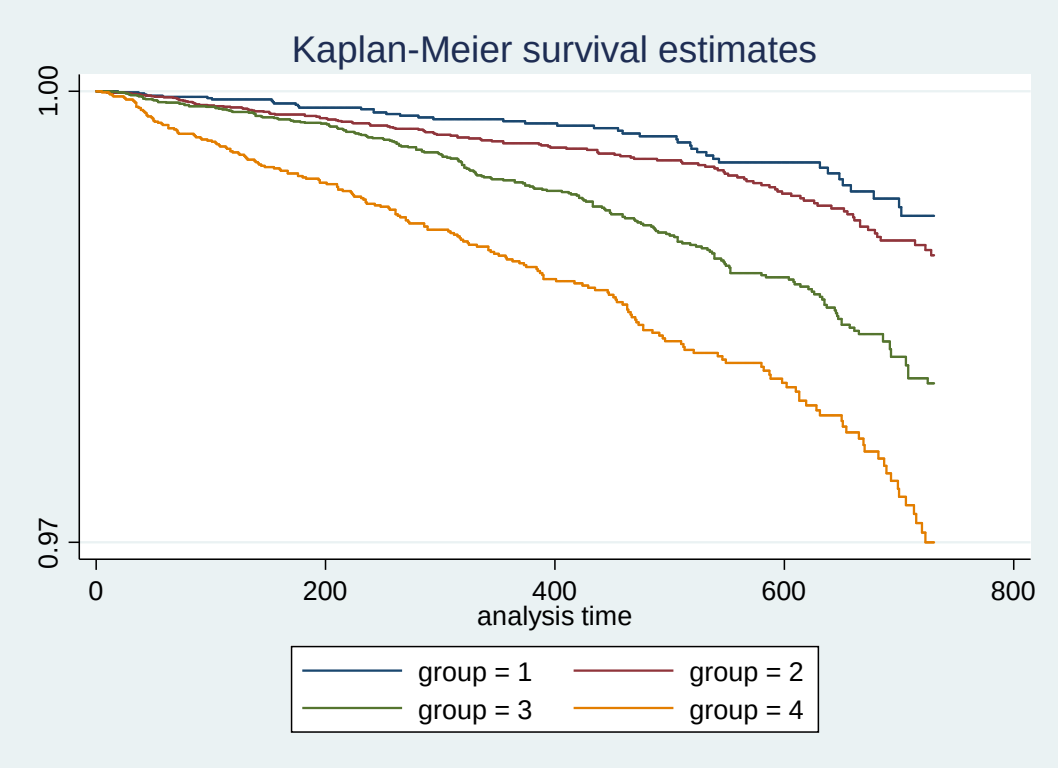


Figure 32: Kaplan Meier curves for 4 risk groups, using the linear predictor which is divided into 4 using Cox's method, for the model derived from a population with negative FOBTs only.

3.6.2 Adjusting for Optimism

The optimism of the model was assessed by calculating Van Houwelingen's heuristic shrinkage which was 0.932 ((191.417- 13)/ 191.417). The linear predictor was then reassessed after applying this shrinkage factor and compared to the original linear predictor this had a mean of 0.140 (SD: 0.421) and range -0.553 to 4.099 (**Figure 33**). The calibration slope after applying the shrunken linear predictor was 1.073. **Figure 34** shows visually how the survival is adjusted after applying shrinkage for a high risk individual with a linear predictor of value 3.039 and a low risk individual with a linear predictor of -0.593. The calibration slope after adjusting for optimism was 1.073.

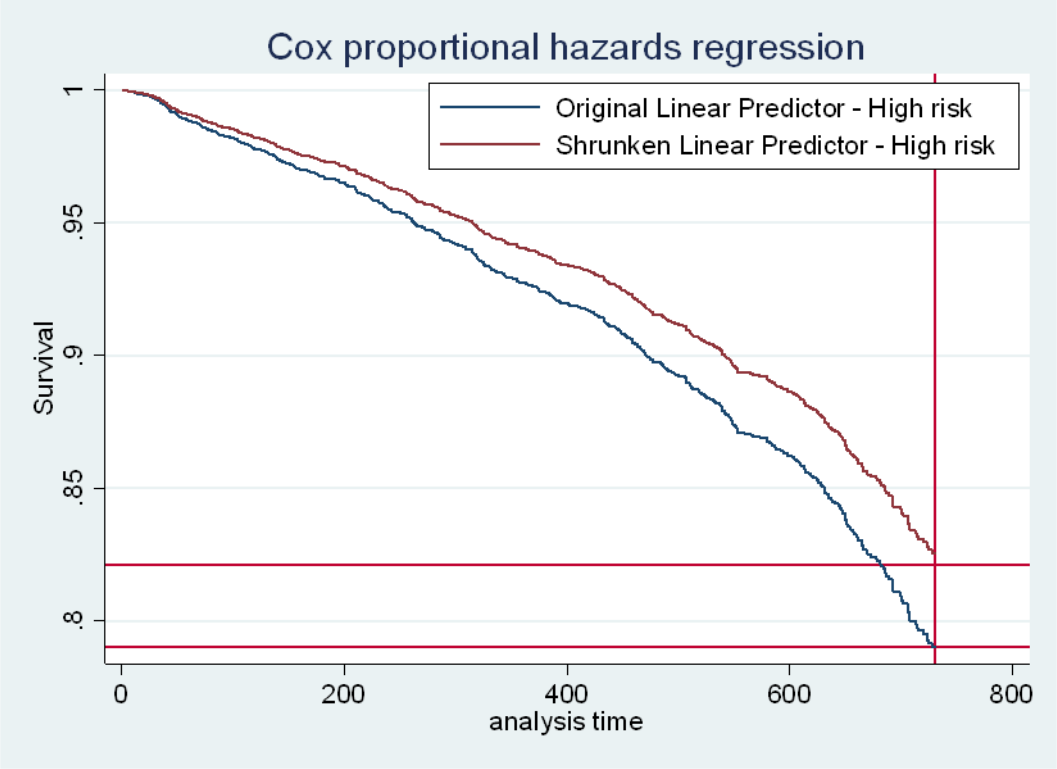


Figure 33: Survival probability plot for a high risk participant using the original linear predictor 3.039 and shrunken linear predictor 2.833 (from the model developed from the sample population with negative FOBTs only)

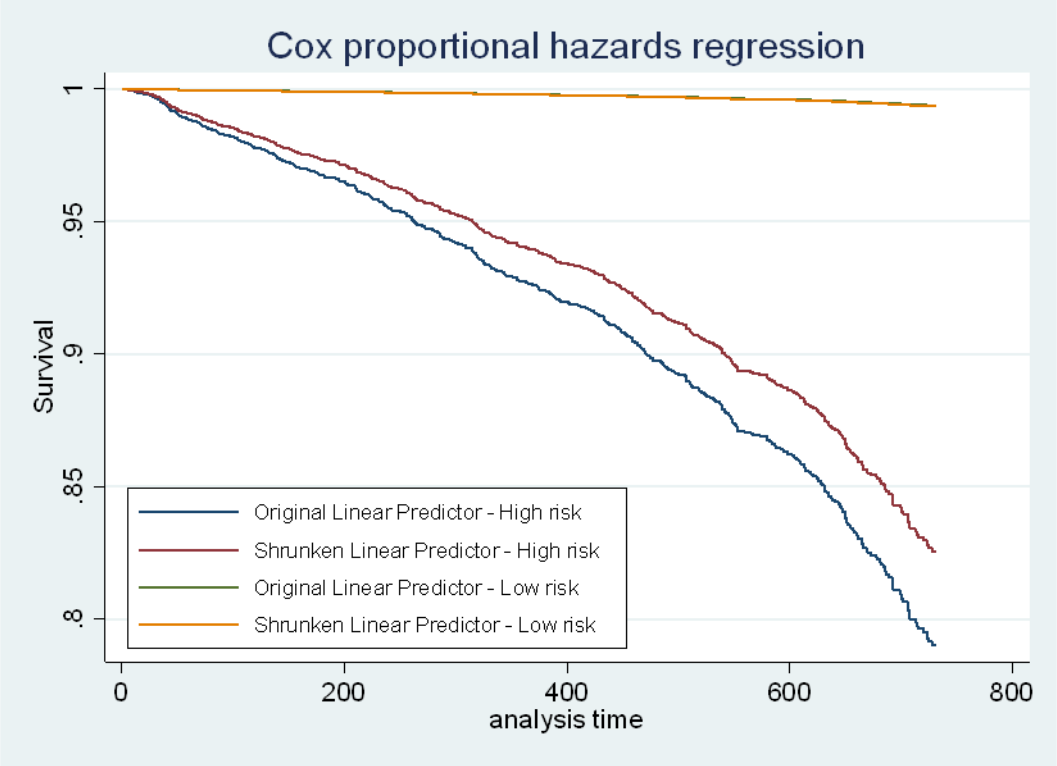


Figure 34: Survival for a high risk individual with a linear predictor of 3.039 which is shrunk to 2.833 and a low risk individual -0.593 which is shrunk to -0.553 (from the model developed from the sample population with negative FOBTs only).

The optimism adjusted values for the C statistic, c-slope, D statistic and R^2 these performance parameters are displayed in **Table 18**. The bootstrapped uniform shrinkage factor (based on the optimism adjusted c-slope value) is slightly higher (0.944) compared to the heuristic shrinkage (0.932).

Statistic	Apparent Performance	Optimism (100 bootstrap replications)	Optimism adjusted performance (apparent minus optimism)
C statistic	0.658	0.008	0.650
c-slope	1.000	0.056	0.944
D statistic	0.906	0.070	0.836
R^2	0.164	0.020	0.144

Table 18: Optimism calculated for the C statistic, c-slope, D statistic and R^2 using 100 bootstrap replications and the corresponding optimism adjusted performance values (from the model developed from the sample population with negative FOBTs only). For bootstrap replications the seed was set as '231398' in Stata.

3.6.3 Predicted Probabilities

This analysis was performed to determine individual risk probabilities from the model and determine the distribution of risk in the sample population based on the predictors in the multivariable model.

The baseline survival for the Cox model was estimated non-parametrically at 2 years as 0.989. After shrinkage the baseline survival was estimated as 0.988. The shrunken baseline hazard was estimated using the heuristic linear predictor and is depicted graphically in **Figure 35**. To generate risk probabilities the heuristic linear predictor and the corresponding shrunken baseline survival were used for the final risk equation. The mean probability of being diagnosed with CRC or polyp within 2 years was 0.015 with standard deviation 0.010 (Range: 0.007, 0.503). This distribution is plotted below in **Figure 36**. The final risk equation for a high risk participant is shown in **Equation 3**.

A Nomogram for this model as an alternative method of presenting the risk equation is provided in **Figure 37**. The Nomogram gives the survival probability and therefore to obtain the event probability this would need to be subtracted from 1.

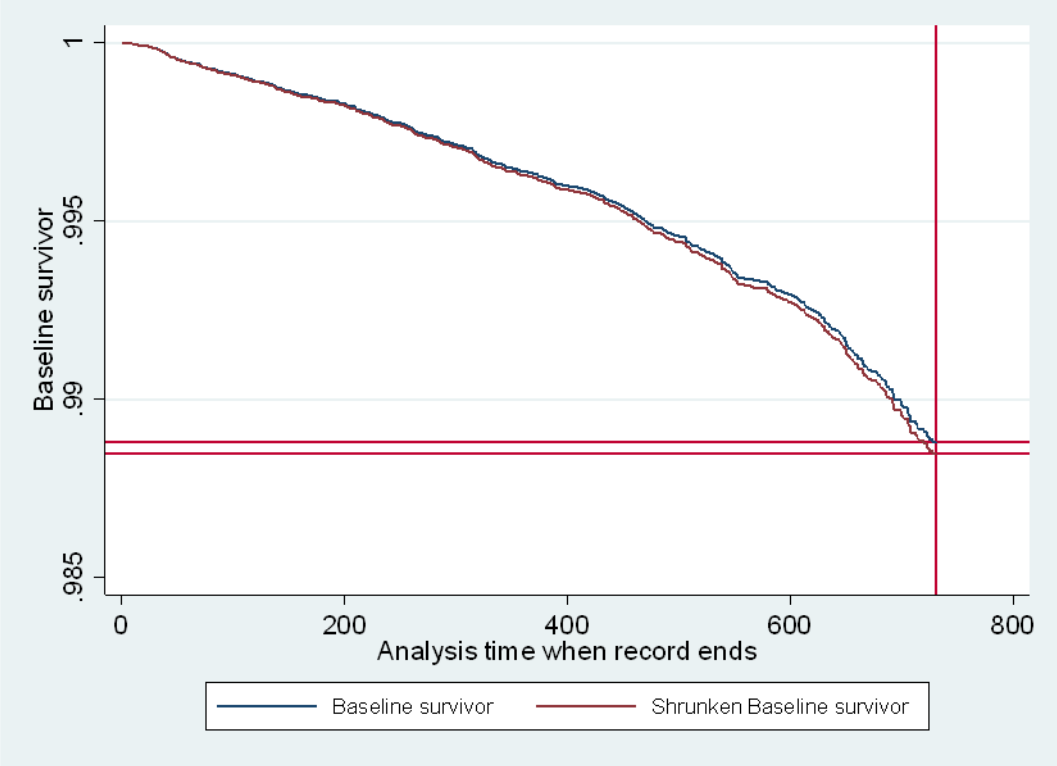


Figure 35: Baseline survivor versus the shrunken baseline survivor derived from the model developed for a population with negative FOBTs only. The shrunken baseline survival at 2 years was estimated by setting the shrunken linear predictor as an offset and predicting the subsequent baseline survival.

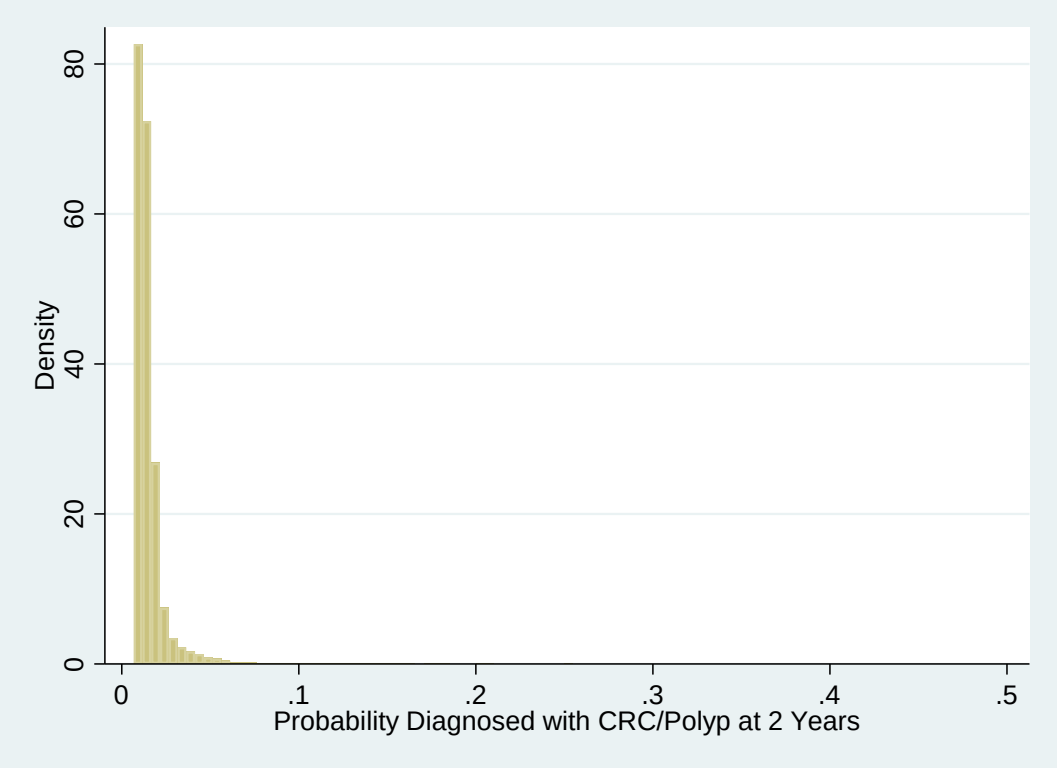


Figure 36: Histogram of the individual probabilities of being diagnosed with colorectal cancer/polyp in a 2 year period for the model derived from a sample population with negative FOBTs only. This model uses the heuristic linear predictor and the corresponding shrunken baseline survival to generate event probabilities.

Survival Probability

$S(2) = S_0(2)^{\exp(LP)}$

Where LP is the linear predictor and $S_0(2)$ is the baseline survival at 2 years.

Event Probability

P = 1 – S(2)

High risk Participant Example:

Survival Probability:

$0.82114774 = 0.9884683^{\exp(2.8325753)}$

Event Probability:

0.17885226 = 1-0.82114774

The probability of being diagnosed with CRC in a 2-year period for a high risk individual is 0.18.

Full Equation:

Survival Probability

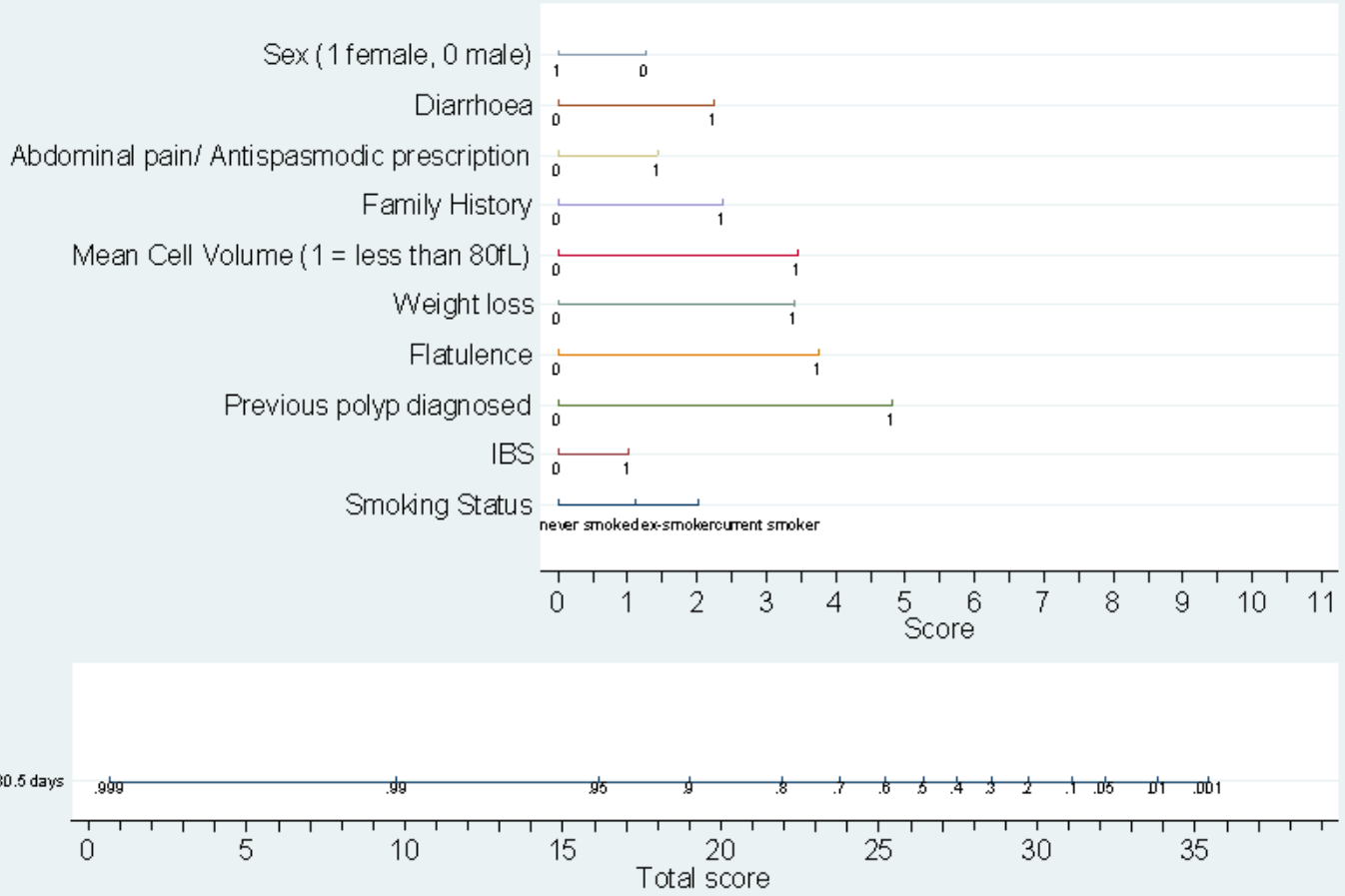
$$S(2) = 0.989^{\exp(0.29x_1 + 0.52x_2 + 0.26x_3 + 1.23x_4 + 0.95x_5 + 0.87x_6 + 0.88x_7 + 0.60x_8 + 0.37x_9 + 0.57x_{10} + (-0.32)x_{11} + 0.03(x_{12} - 66.97) + 0.79x_{13})}$$

0.988 = the baseline survival at 2 years $S_0(2)$
Where S(2) is the survival probability at 2 years (probability of not being diagnosed with colorectal cancer/polyps)

Event Probability
P = 1 – S(2)
Where P is the probability of colorectal cancer/polyp being diagnosed within 2 years of the latest FOBT date; x_1 ex-smoker; x_2 current smoker; x_3 IBS; x_4 previous polyps; x_5 flatulence; x_6 weight loss; x_7 MCV <80fL; x_8 family history of gastrointestinal cancer; x_9 abdominal pain; x_{10} diarrhoea; x_{11} sex; x_{12} age at FOBT; x_{13} change in bowel habit.

Equation 3: Final risk equation for the model derived from participants with negative FOBT only, using the shrunken baseline survival and the shrunken linear predictor values to correct for optimism.

Nomogram


Sex (1 female, 0 male)
Diarrhoea
Abdominal pain/ Antispasmodic prescription
Family History
Mean Cell Volume (1 = less than 80fL)
Weight loss
Flatulence
Previous polyp diagnosed
IBS
Smoking Status
never smoked ex-smoker current smoker
0 1 2 3 4 5 6 7 8 9 10 11
Score
730.5 days
.999 .99 .95 .9 .8 .7 .6 .5 .4 .3 .2 .1 .05 .01 .001
0 5 10 15 20 25 30 35
Total score

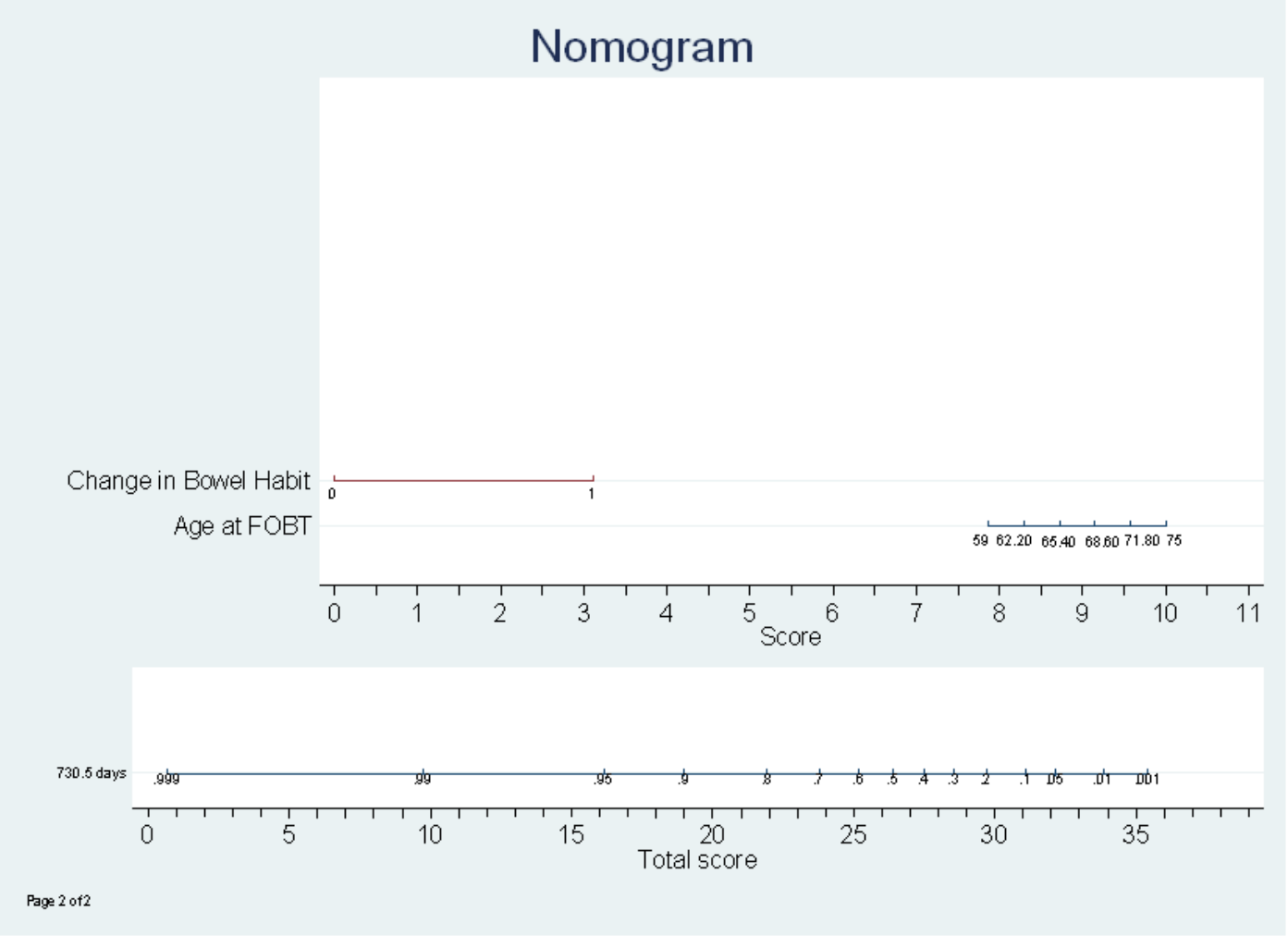
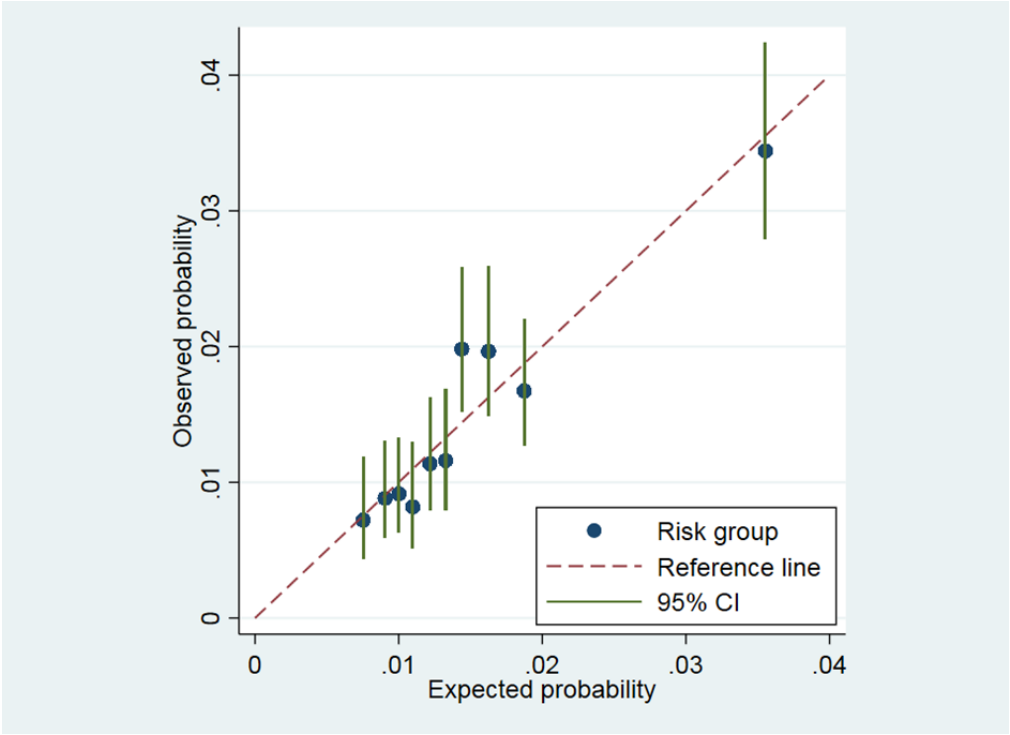


Figure 37: Nomogram for the final Cox Regression model for participants with a negative FOBT only which gives the colorectal cancer/polyp free survival probability. To obtain the event probability subtract the survival probability from 1.

3.6.4 Calibration

A calibration curve for the multivariable model adjusted for optimism is presented below for deciles of risk (**Figure 38)**. Compared to the multivariable model including the FOBT result, the spacing between groups was more even but there was still a higher risk group, possibly due to the presence or absence of a strong predictor (e.g. previous polyps). In addition, most of the groups lie close to the line of equality, indicating good calibration.



Risk Group	Calibration Expected Probability	Calibration Observed Probability	Calibration observed lower bound	Calibration observed upper bound
1	0.008	0.007	0.012	0.004
2	0.009	0.009	0.013	0.006
3	0.010	0.009	0.013	0.006
4	0.011	0.008	0.013	0.005
5	0.012	0.011	0.016	0.008
6	0.013	0.012	0.017	0.008
7	0.014	0.020	0.026	0.015
8	0.016	0.020	0.026	0.015
9	0.019	0.017	0.022	0.013
10	0.036	0.034	0.042	0.028

Figure 38: Calibration plot of observed probability versus expected probability using the multivariable model of participants with negative FOBTs only. The corresponding risk groups for each decile of probability are presented in the table below the figure.

3.6.5 Cox Regression Diagnostics

As with the model including both negative and positive FOBT results, Schoenfeld residuals were examined to test the proportional hazards assumption of the Cox Regression model (**Appendix 7** for full results). Predictors with a p value of less than 0.05 included; smoking status (previous smoker) and age at FOBT. In addition, the global test of proportional hazards had a p value of 0.016. This statistical test is dependent on sample size and so the larger sample size means significance is more likely to be detected. Scaled Schoenfeld residual plots (**Figure 39**) along with the log-log plots (**Figure 40**) were plotted for these two significant variables. The Schoenfeld residual plots have a roughly straight line but the log-log plots suggest they potentially violate the proportional hazards assumption.

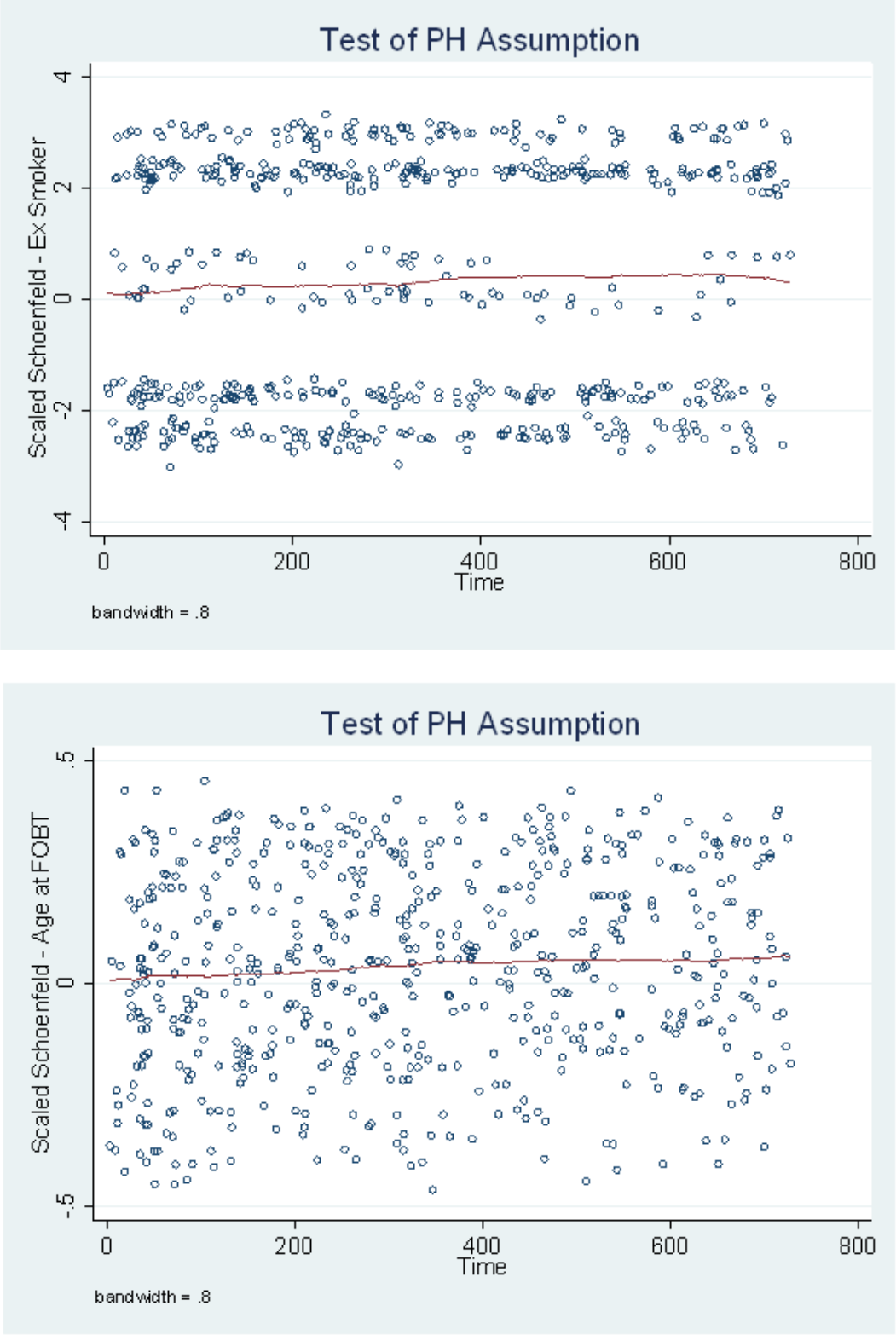
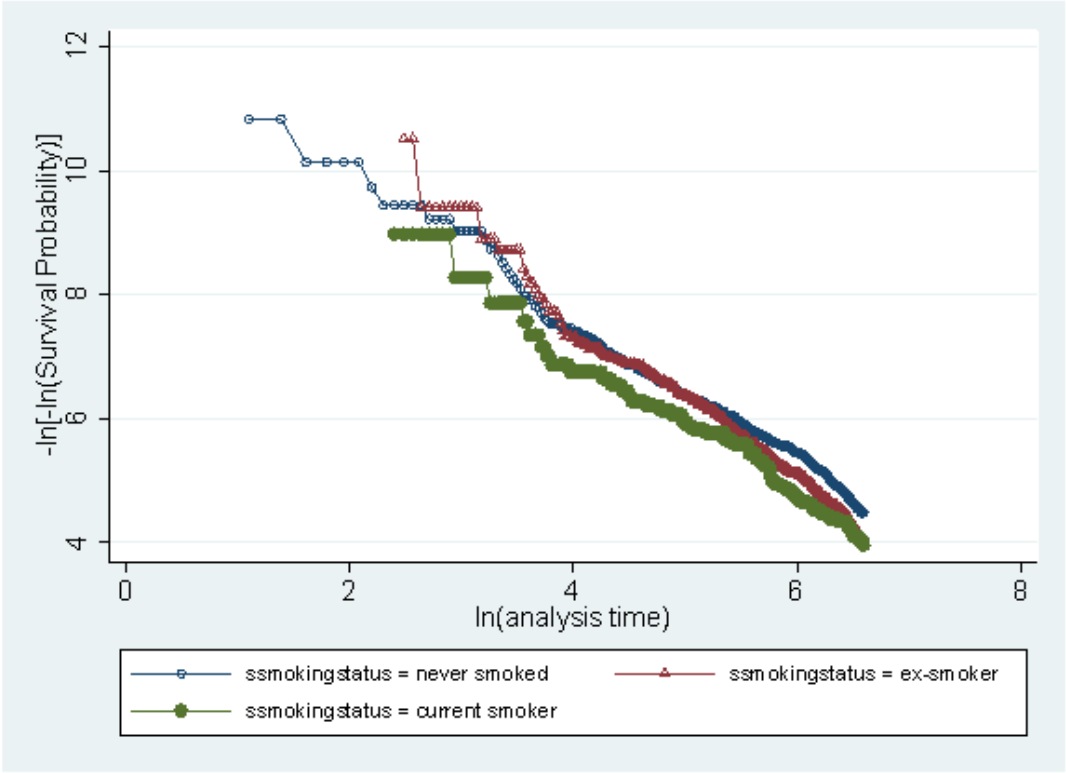


Figure 39: Schoenfeld residual plots for variables which had a p value of <0.05 when testing the proportional hazards assumption in the multivariable model with negative FOBTs only. These variables included: ex-smoker (1st plot), age at FOBT (2nd plot).



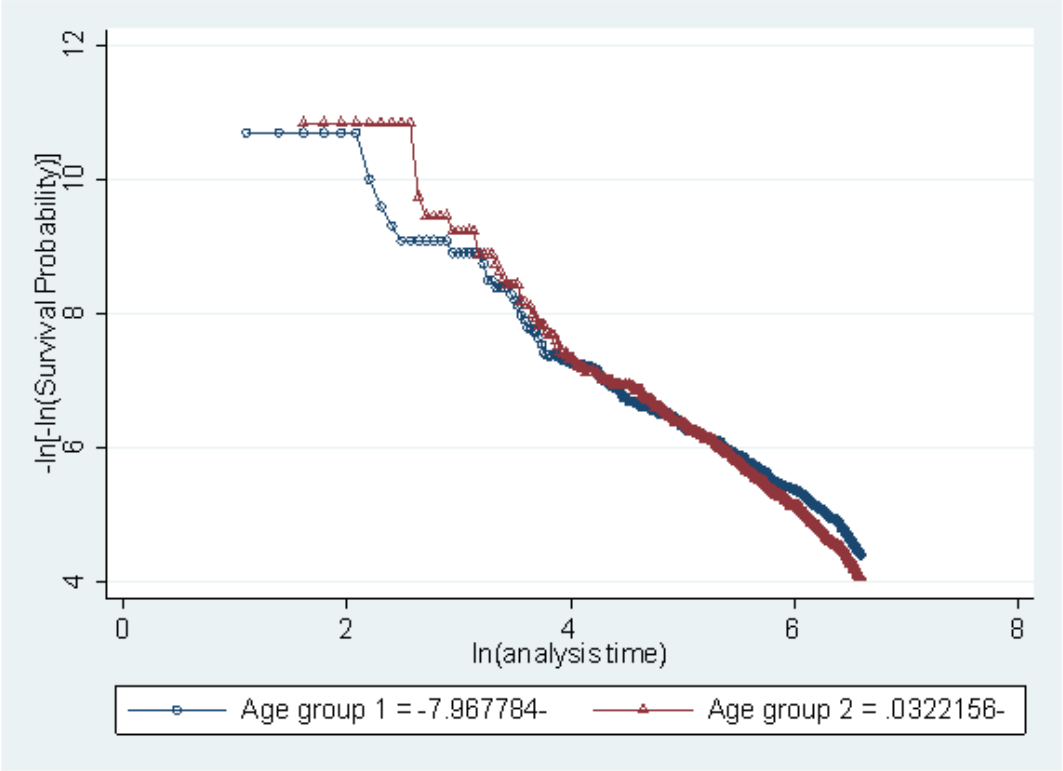


Figure 40: Log-log plots to test the Cox proportionality assumption for smoking status (1st plot), and Age group (2nd plot - which was split into 2 equally sized groups) in the multivariable Cox Regression Model with negative FOBTs only.

The overall model fit was assessed using Cox-Snell residuals (**Figure 41**). The cumulative hazard function deviates from the line at the tail end but roughly follows an exponential distribution with a hazard rate of one. Parametric survival models were subsequently investigated to determine whether these more flexible models have a better fit.

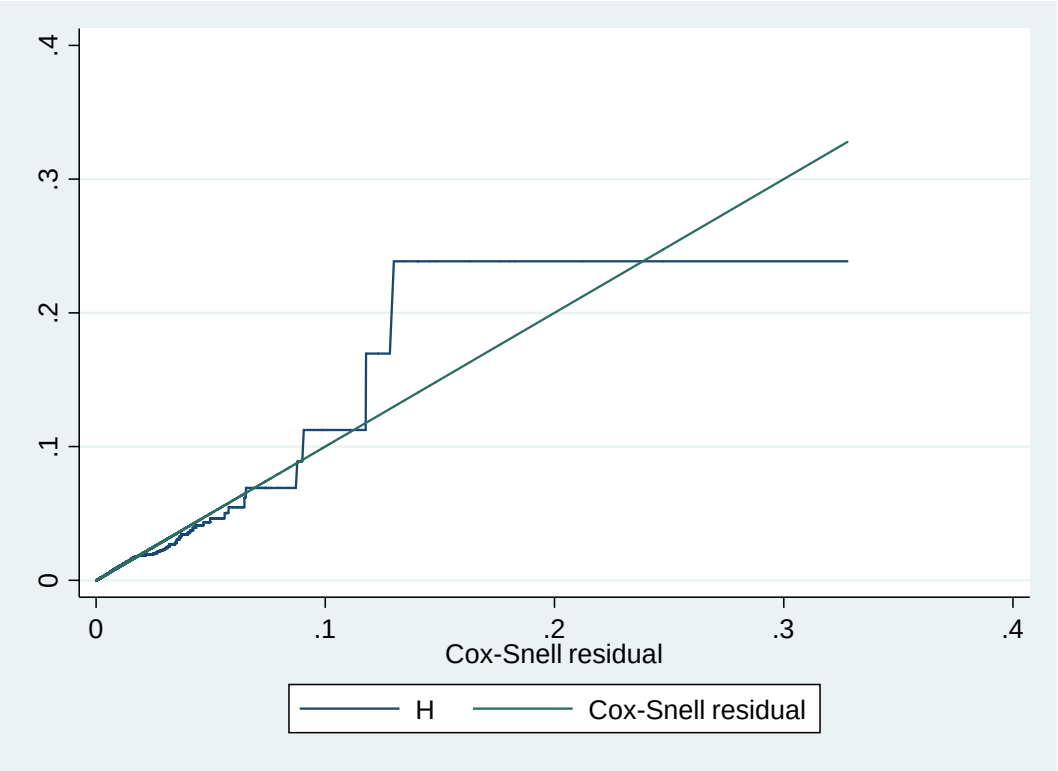


Figure 41: Assessment of overall model fit of the Cox Regression model (negative FOBTs population) using Cox-Snell residuals and plotting the Nelson-Aalen cumulative hazard function against Cox-Snell residuals. For a good model fit, the cumulative hazard function should follow the Cox-Snell residuals.

3.6.6 Parametric Survival Models

Parametric models were investigated as an extension to Cox Regression to determine whether these types of model gave a better fit to the data.

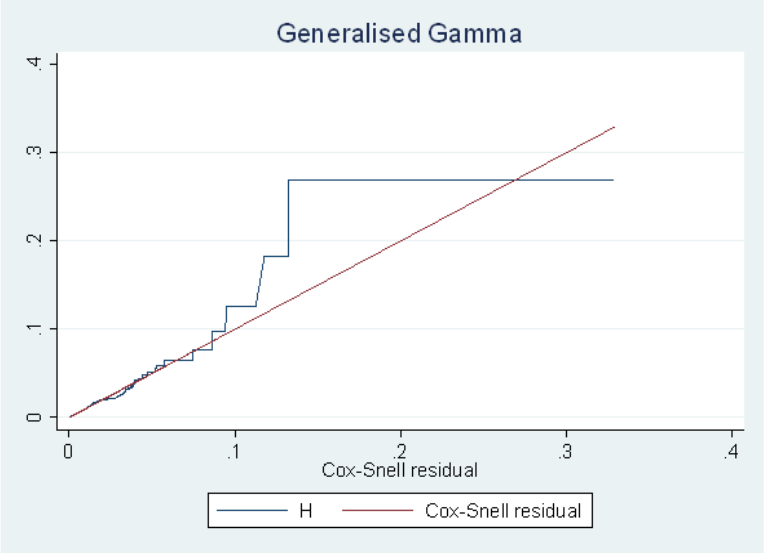
The AIC was used to compare the different parametric survival models (**Table 19**). The smallest AIC was achieved with the Gompertz model. Gompertz models are usually used for growth data where the shape of the hazard distribution is always increasing.[40] The suitability of this model to the data may be because the number of events appears to increase after year two when analysing the KM survival probability curve.

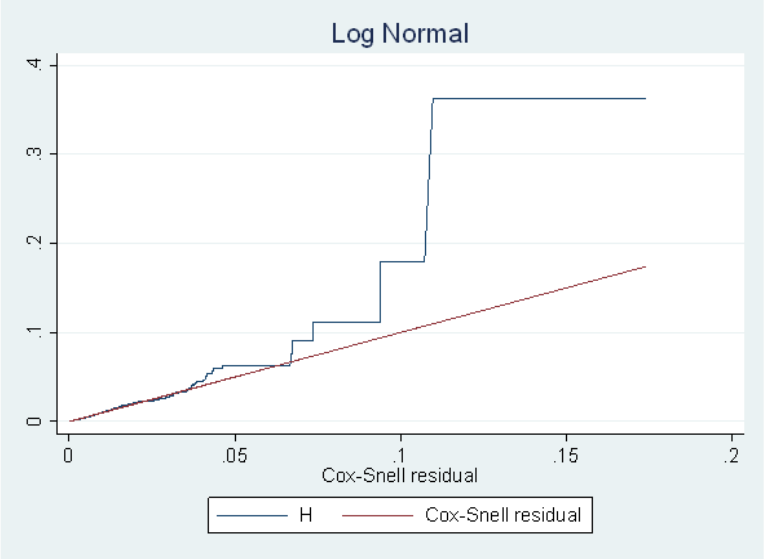
Model	Observations	Log likelihood null	Log likelihood model	Degrees of freedom	AIC	BIC
Exponential	587	-3866.88	-3771.922	14	7571.844	7633.094
Weibull	587	-3840.105	-3744.628	15	7519.257	7584.882
Gompertz	587	-3829.577	-3733.873	15	7497.746	7563.371
Lognormal	587	-3851.828	-3754.748	15	7539.495	7605.121
Loglogistic	587	-3840.337	-3744.852	15	7519.705	7585.33
Generalised Gamma	587	-3839.875	-3744.424	16	7520.848	7590.848

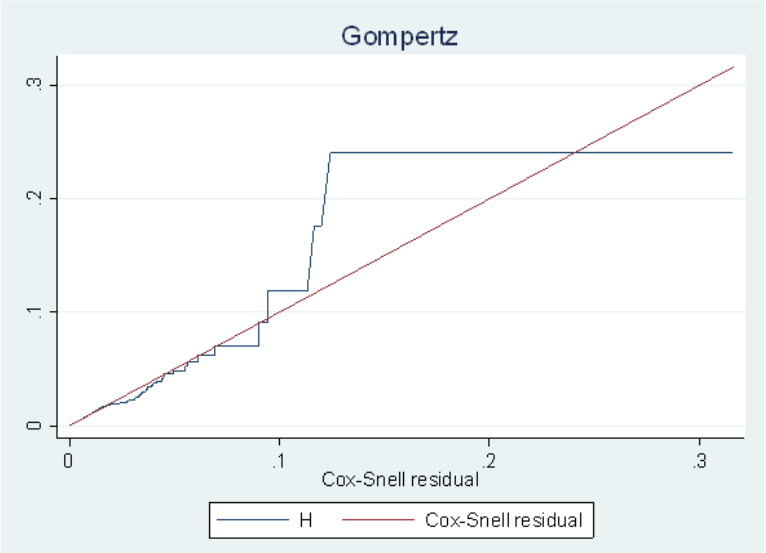
Table 19: Model parameter comparisons for the parametric models derived from a sample population with negative FOBTs only.

Overall model fit of the different parametric models was assessed using Cox-Snell residuals (**Figure 42**). Visually, the Gompertz and generalised gamma model have the better fit to the data compared to the other parametric models. The Nelson Aalen cumulative hazard plots had a very similar fit to the data for all the models (**Figure 43**). Finally, the Gompertz Kaplan Meier function plot had the best fit between the models (**Figure 44**).

The coefficients for the Gompertz model are reported in **Table 20**.


Generalised Gamma
0
.1
.2
.3
.4
Cox-Snell residual
H
Cox-Snell residual


Log Normal
0
.05
.1
.15
.2
Cox-Snell residual
H
Cox-Snell residual


Gompertz
0
.1
.2
.3
Cox-Snell residual
H
Cox-Snell residual

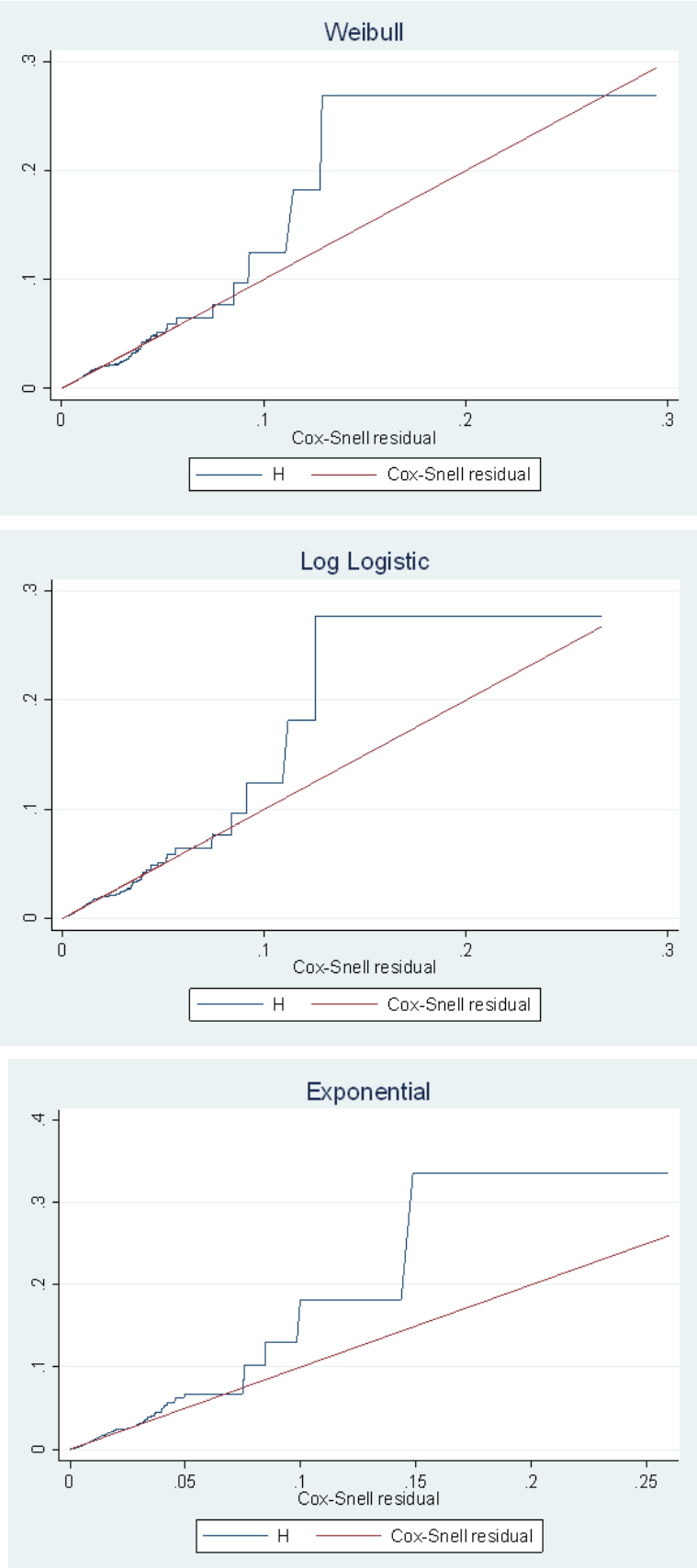
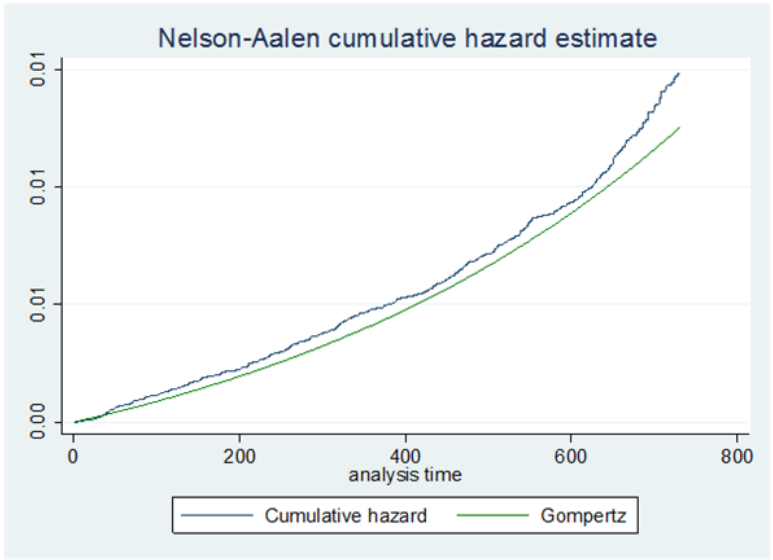
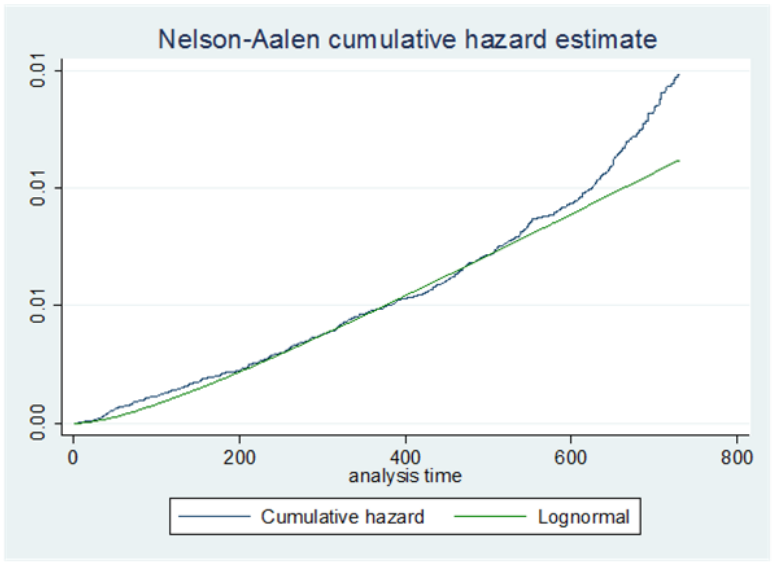
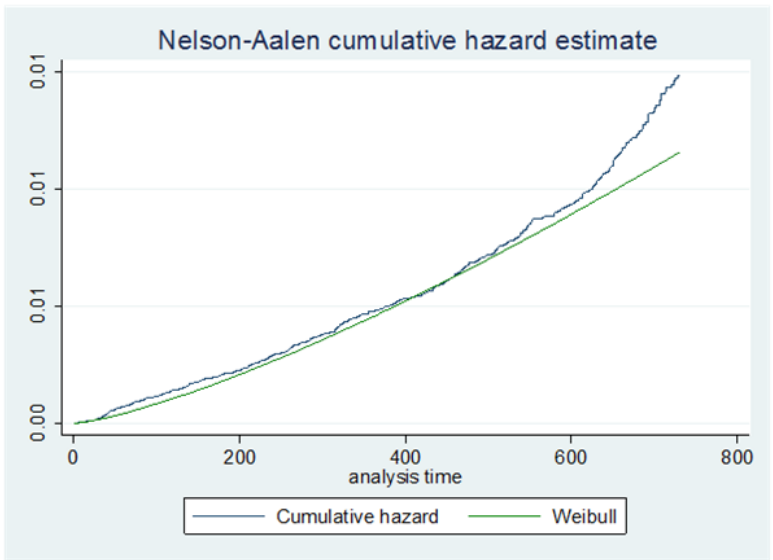


Figure 42: Cox Snell Residuals plotted for all considered parametric models, derived from a sample population with negative FOBTs only, to assess model fit.


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0
200
400
600
800
analysis time
Cumulative hazard
Gompertz


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0
200
400
600
800
analysis time
Cumulative hazard
Lognormal


Nelson-Aalen cumulative hazard estimate
0.00
0.01
0.01
0.01
0
200
400
600
800
analysis time
Cumulative hazard
Weibull

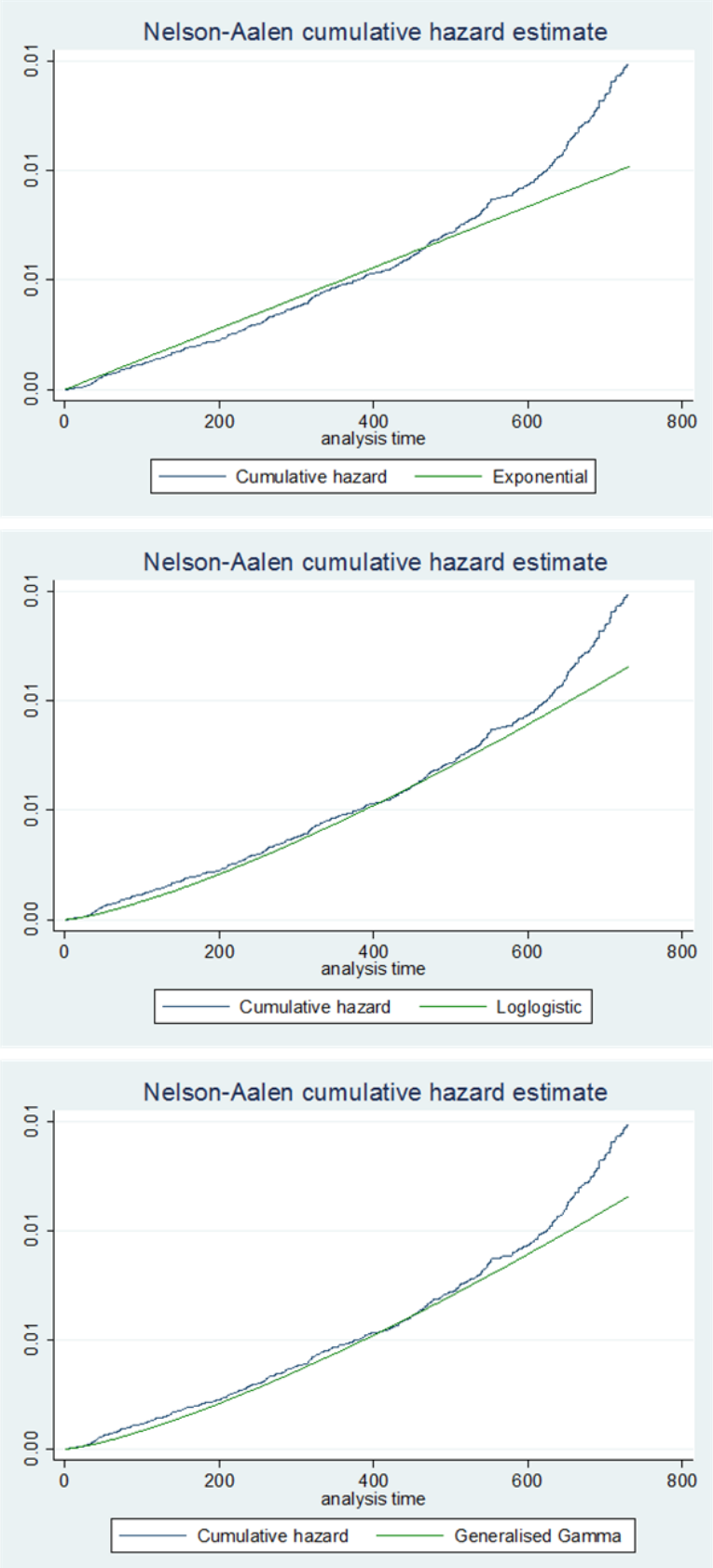
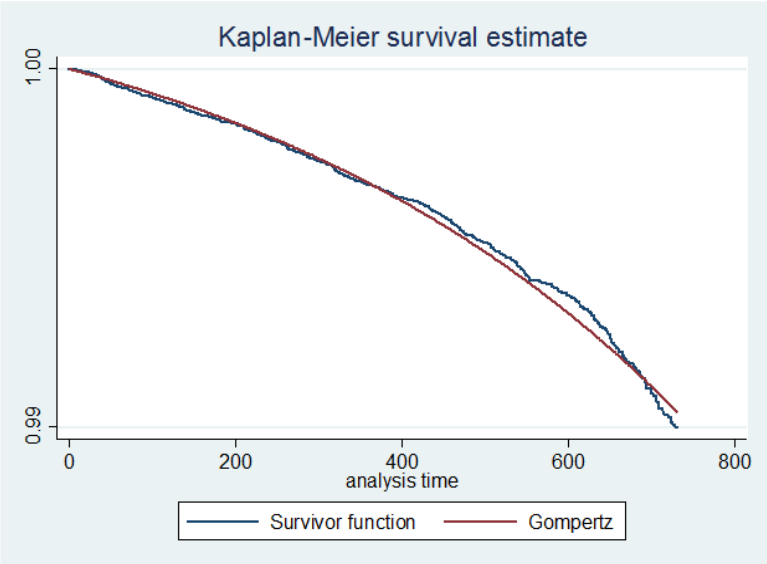
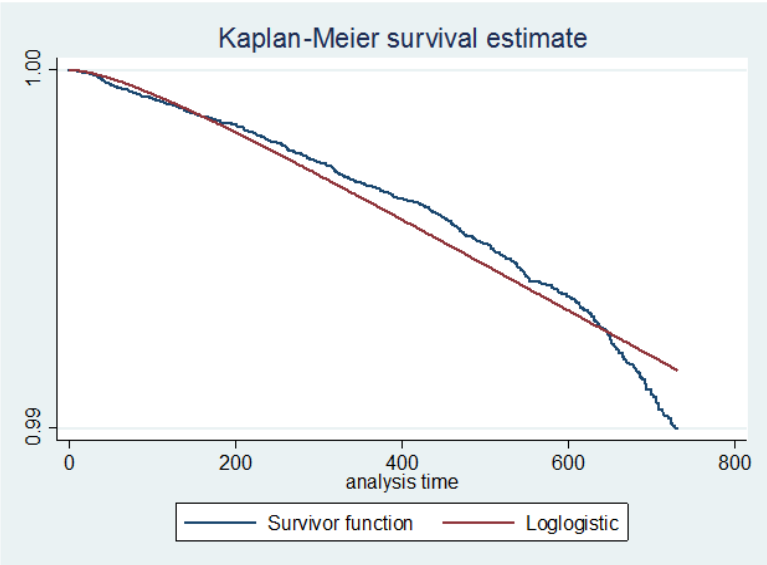
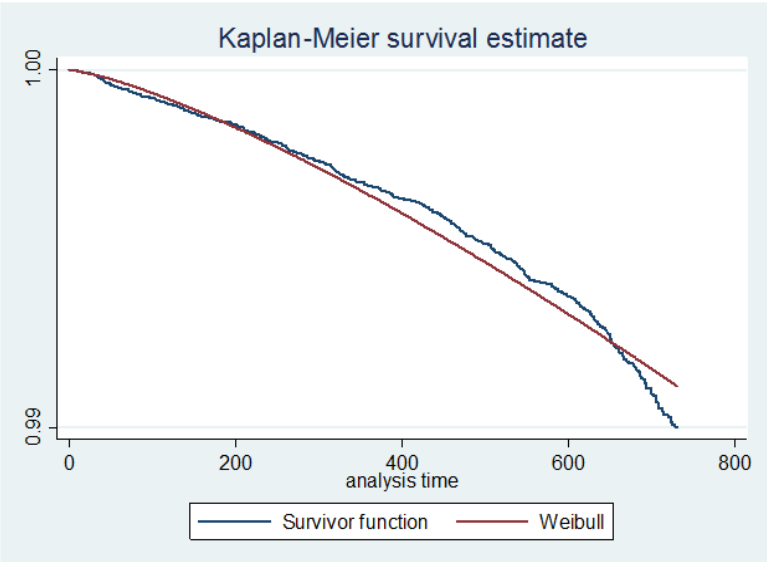


Figure 43: Nelson-Aalen cumulative hazard plots for all considered parametric models, derived from a sample population with negative FOBTs only, to assess model fit.


Kaplan-Meier survival estimate
1.00
0.99
0
200
400
600
800
analysis time
Survivor function
Gompertz


Kaplan-Meier survival estimate
1.00
0.99
0
200
400
600
800
analysis time
Survivor function
Loglogistic


Kaplan-Meier survival estimate
1.00
0.99
0
200
400
600
800
analysis time
Survivor function
Weibull

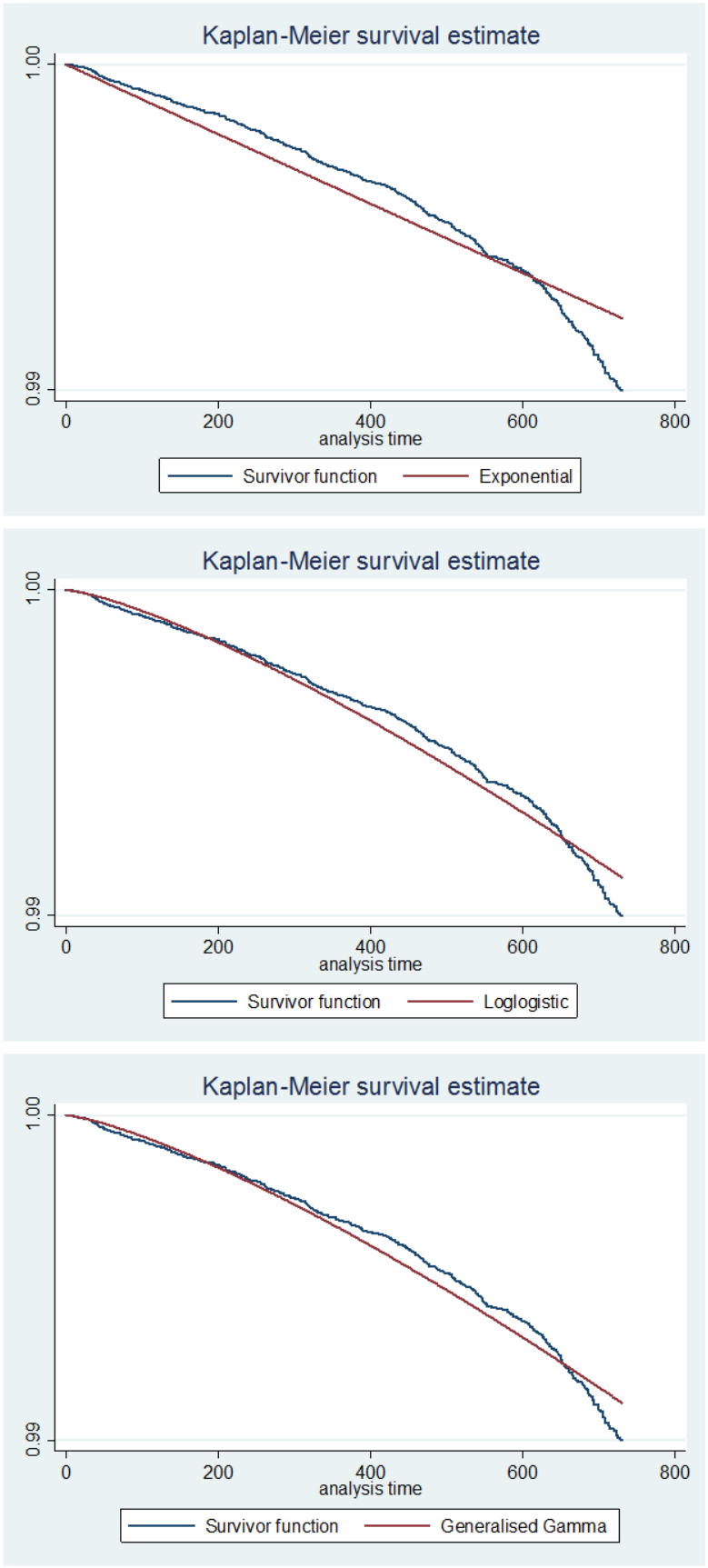


Figure 44: Kaplan-Meier function graphs for all considered parametric models, derived from a sample population with negative FOBTs only, to assess model fit

3.6.7 Model Performance measures for the best fitting parametric models

Based on the AIC value which allows assessment of the relative fit of all the parametric models and the plots presented in the previous section, the Gompertz model had the best fit to the data. **Table 20** shows the coefficients and the discrimination of the model in comparison to the Cox Regression model. Both the hazard ratios and coefficients are presented to aid comparison. The Gompertz and Cox regression models have very similar coefficients and the same discrimination as reflected by Harrell's C statistic (95% CI: 0.658, 0.633, 0.683). The Gompertz model has slightly less optimism as shown by the C-slope 0.950 versus 0.944 but the differences between the two models are negligible.

Variable	Gompertz coefficients (PH)	Gompertz Hazard Ratios (PH)	Cox coefficients	Cox Hazard Ratios
Smoking Status:				
ex-smoker	0.286	1.331	0.285	1.330
current smoker	0.521	1.684	0.516	1.676
IBS	0.258	1.295	0.258	1.294
Previous Polyps Diagnosed	1.225	3.4055	1.225	3.403
Flatulence Symptom Recorded	0.959	2.610	0.953	2.594
Weight loss	0.864	2.373	0.867	2.379
MCV <80fL	0.877	2.403	0.877	2.403
Family History of Gastrointestinal Cancer	0.604	1.829	0.603	1.828
Abdominal pain/antispasmodic prescription recorded	0.364	1.439	0.365	1.440
Diarrhoea symptom	0.575	1.776	0.572	1.772
Sex	-0.325	0.722	-0.323	0.724
Age at FOBT	0.033	1.034	0.034	1.035
Change in bowel habit symptom	0.788	2.199	0.793	2.209
Constant	-11.878	6.94E-06	-	-
Ancillary parameter	0.002	0.002	-	-
Log likelihood		-3733.873		-6233.842[a]
AIC		7497.746		12493.680[a]
BIC (n=587)		7563.371		12550.560
Harrell's C Statistic (95%CI)		0.658 (0.633, 0.683)		0.658 (0.633, 0.683)
R^2		0.164		0.164
D Statistic		0.906		0.906
Adjusted R^2 (Bootstrap CI 100 replications)		0.151 (0.113, 0.192)		0.151 (0.113, 0.191)
Optimism adjusted Calibration Slope (also shrinkage factor for linear predictor)		0.950		0.944
Optimism Harrell's C Statistic		0.650		0.650
Optimism adjusted D statistic		0.852		0.836
Optimism adjusted R2		0.147		0.144

[a] The log likelihood/AIC/BIC from the Cox model is not comparable to parametric models since it uses partial likelihood whereas the other models use full maximum likelihood.

Table 20: Comparison of the best fitting parametric model compared to the Cox model for a sample population with negative FOBT results. Model coefficients, model constants, ancillary parameters, AIC, BIC, R^2, D statistic and optimism adjusted performance metrics are presented for comparison. The R^2 used in this instance is Royston and Sauerbrei's (2004) R^2_D measure of explained variation for survival models based on their index of discrimination (D. [73] The adjusted R^2 measure also considers the number of covariates in the model. For non-proportional hazards models R^2 for explained variation is not interpretable but can be used as an index of determination.[74]

A calibration plot for the Gompertz model is presented in **Figure 45** below. The decile groups show a similar pattern along the 45 degree line to the Cox regression calibration plot in **Figure 38**. These investigations suggest that the Gompertz model has a very similar fit to the Cox model. The advantages of the parametric models over their semi-parametric

counterpart is that provided the underlying assumptions are true, they can give more precise parameter estimates. Furthermore, parametric models give greater flexibility with post-estimation since they can provide predicted survival and hazard functions, median survival times as well as predicted probabilties at different timepoints without a need to estimate the baseline hazard using non-parametric methods.

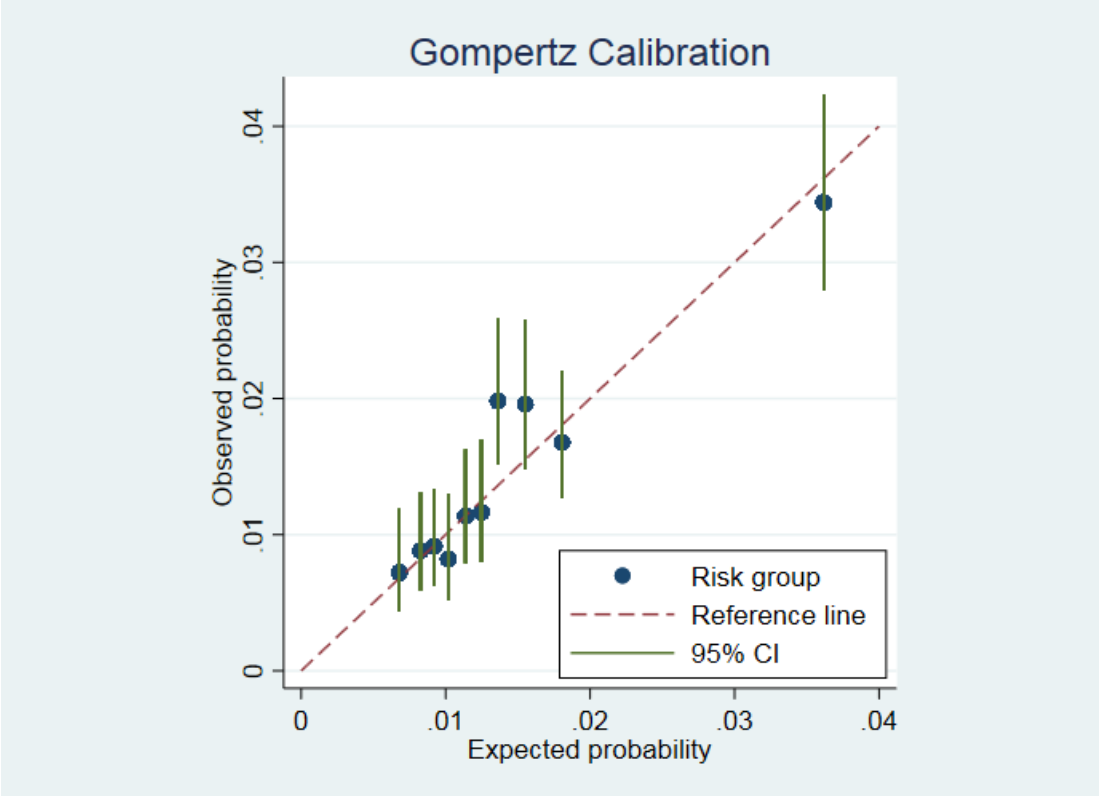


Figure 45: Calibration plot of observed probability versus expected probability using the Gompertz parametric model for a sample population with negative FOBTs only.

4.0 DISCUSSION

4.1 Statement of principal findings

This chapter used the THIN database of anonymised GP records to investigate the availability of GP data for key predictors of colorectal cancer in the screening population and used these data to determine whether additional predictor information can be used to make more accurate screening referral decisions by developing two multivariable prediction models. Descriptive analyses included determining diagnostic accuracy results of the FOBT and survival analyses investigating time to diagnosis (colorectal cancer free survival) and time to all cause death (overall survival) stratifying by FOBT result, sex and 2 by 2 category (TP, TN, FP, FN).

The test positivity (2.18% for the overall cohort), sensitivity (52.8%) and specificity (96.3%) of the guaiac FOBT were calculated for the derived screening population and were found to be similar to reported values in the literature.[76] [6]

Survival analysis based on time to diagnosis (colorectal cancer free survival) and time to death (overall survival) was investigated for the screening cohort stratified by FOBT result and sex. Patients with a positive FOBT had a much sharper rate of decline in survival (increase in diagnosis) compared with a negative test result (significant log-rank test for equality $p<0.001$). Males have a significant reduction in survival (increase in colorectal cancer/polyp diagnoses) over time compared to females for the two year follow up period. For time to all cause death, people with a positive FOBT were at increased risk for first 1250 days, for the last 1000 days those with negative FOBT results appear to be at greatest risk.

In addition, Kaplan Meier curves were plotted for TP, TN, FP, FN results for time to diagnosis and time to all cause death. For time to all cause death, the FNs end up doing worse later on than the TPs at around 1000 days, this could be due to the effects of late diagnosis for FNs as they are less likely to have had diagnostic follow up/further investigations compared to TPs. In addition, FNs could have more aggressive cancers that develop after screening and TPs present at an earlier stage. FPs also appear to have less diagnoses than TNs possibly due to having some sort of diagnostic investigation putting them at lower risk of CRC compared to TNs.

Symptoms and diagnoses as binary parameters were 100% reported if a patient had consulted their GP. Lab measurements including platelet count, MCV and haemoglobin concentration were recorded for about 45% of those adequately screened with a positive

or negative FOBT result. Ferritin was less well reported at around 8.59%. Lifestyle factors such as smoking status were extremely well recorded (99.42%) and alcohol consumption was also fairly well recorded at 78.00%.

Univariable Cox Regression revealed that screening history variables such as previous positive FOBT results and previous polyps diagnosed had the largest observed hazard ratios at 5.032 (CI:4.184-6.052) and 3.182 (CI:2.503-3.883) respectively. Rectal bleeding or melaena was the symptom with the highest HR 3.118 (2.503-3.883). Age, BMI and Alcohol consumption were all significant and females were at lower risk of colorectal cancer/polyp diagnosis than males (HR 0.656: CI 0.609-0.706).

After these investigations, a multivariable Cox Regression model was built for the screening population with a positive or negative FOBT result (n=98,303). This model combined the FOBT result with other risk factors to identify additional predictors which could be included in a risk based model to make more accurate screening referral decisions. The final model included the following variables: FOBT result, smoking status, whether a patient had a diagnosis of Crohn's disease, previous polyps diagnosed, flatulence, MCV of <80fL compared to a MCV of ≥80fL, alcohol consumption in units per week, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit. Significant interactions at the 0.05 p value level included FOBT result and age and MCV and age.

The optimism of the model was assessed by calculating Van Houwelingen's heuristic shrinkage which was 0.995 and was applied to the linear predictor. The calibration slope after applying the shrunken linear predictor was 1.005. The shrunken linear predictor had a mean of 0.135 (SD: 0.685) and range -1.365 to 5.728. Internal validation using 100 bootstrap replications was used to adjust the performance parameters for optimism giving; a C statistic of 0.850, c-slope of 0.991, D statistic 2.298 and R^2 of 0.558. To generate risk probabilities both the heuristic linear predictor and the corresponding shrunken baseline survival were used for the final risk equation. The calibration plot shows the 10^{th} group was far removed, most likely due to whether an individual has either a positive or negative FOBT.

Overall model fit was assessed using Cox-Snell residuals and revealed that a semi-parametric model may not provide the best fit to the data as the cumulative hazard function did not appear to have an exponential distribution with a hazard rate of one. The best fitting parametric survival models based on the AIC, cumulative hazard plots, Kaplan

Meier function plots and Cox-Snell residuals was the generalised gamma model. This model had similar discrimination (0.859 (95% CI: 0.845, 0.872)) to the Cox model (0.854 (95% CI: 0.841, 0.868) and the observed versus predicted risk groups were slightly closer to the 45 degree line on the calibration plots but were also comparable to the Cox model. A Wald test for the hypothesis of the kappa ancillary parameter of the gamma model being equal to 1 was significant suggesting a potentially good fit for the Weibull model also (C-statistic 0.854 (95% CI: 0.841, 0.868)).

Cox regression was then used to investigate additional predictors which could be used in a screening population with negative results (n=95,792) for screening referral decisions. The model was built using the same methods described above. Predictors which could be used to determine the risk of whether an individual has colorectal cancer/polyps included; smoking status, whether a patient had an IBS diagnosis, previous polyps diagnosed, flatulence, weight loss, MCV of <80fL compared to a MCV of ≥80fL, family history of gastrointestinal cancer, abdominal pain/antispasmodic prescription, diarrhoea, sex, age at FOBT and change in bowel habit.

Van Houwelingen's heuristic shrinkage which was 0.932 was applied to the linear predictor. The shrunken linear predictor had a mean of 0.140 (SD: 0.421) and range -0.553 to 4.099. Internal validation using 100 bootstrap replications was used to adjust the performance parameters for optimism giving; a C statistic of 0.650, c-slope of 0.944, D statistic 0.836 and R^2 of 0.144. Absolute risk probabilities were produced using the shrunken linear predictor and baseline survival. The calibration plot produced showed better separation than the previous model due to the removal of FOBT.

Parametric survival models were also investigated, with the Gompertz model having the best fit based on the AIC, cumulative hazard plots, Kaplan Meier function plots and Cox-Snell residuals. The discrimination of the model was the same as the Cox regression model (C-statistic 0.658 (95% CI: 0.633, 0.683)) with a similar calibration as shown in the calibration plots. This suggests that a Cox regression model has the same performance in this sample population.

4.2 Strengths and weaknesses of the study

A major strength of the study is the sample size used for assessing test accuracy measures, survival analysis and developing the multivariable prediction models. The screening population derived from THIN can be considered representative of the average risk screening population when assessing demographic factors and test accuracy measures.

The AEB date was derived as part of this THIN work and defines the start date at which GP practices started to receive electronic notifications from the NHS BCSP. This date acts as a level of quality assurance since before this date, paper records would have used and there would have been a bias to recording positive FOBT results. The electronic notifications use the same system as Pathlinks for laboratory test results. The AMR date was also identified for each practice as another level of quality assurance. Before this date each practice may not have routinely recorded patient deaths and de-registrations. Patflags (patient flags assigned by THIN) were used to check the integrity of the data and ensure quality.

The QOF is an incentive programme for GP surgeries. Achievement points have been introduced for various conditions which has helped to improve recording of certain parameters in GP records. For instance, practices regularly record smoking status as part of the QOF. For Additional Health Data (AHD) variables (laboratory parameters) an external dataset was used to define the range of expected values to remove potential outliers and ensure the distribution of results were similar in the THIN dataset compared to routine lab measurements. This is reported further in **Chapter 6**.

The extensive number of CRC predictors examined give an indication of what might be important for a screening population. The methods used to derive these data using Read codes and other strategies were thorough and subjected to review by two people, improving the reliability of the data extracted and giving confidence to the associations identified. This is also reported further in **Chapter 6**.

Model development used the latest recommended methods in this field to ensure a reproducible, generalizable and well discriminating model was produced. For instance, the model was subjected to internal validation, used an appropriate sample size with enough events, was corrected for optimism, optimism adjusted performance measures were reported using bootstrapping and the model coefficients along with the baseline survival at 2 years were supplied. Model development followed the TRIPOD guidelines and applied the

recommendations from the systematic review in **Chapter 2** in terms of improving reporting of risk prediction studies.

Limitations include those which relate to the study and those which relate to the GP record database. The sample size for model development was reduced based on using complete variable data. The factors which limited the sample size for this investigation were the laboratory results (Hb concentration, platelet count and MCV). The cancer/polyp detection rate for those with a laboratory record (for all three results) was around 1.19% and those without 0.83% (Pearson's chi-squared p=<0.001). Multiple imputation was considered, however the missing data mechanism for the majority of these predictors would be 'Missing not at random' (MNAR). Individuals who have a blood test result for example are more likely to have this investigation based on suggestive symptoms of a particular underlying disease. Furthermore, a study found that weight and blood pressure may be MAR (Missing at Random) but for lifestyle factors smoking and alcohol consumption, these were not MAR.[80] This may have a corresponding effect on model parameter estimates.

When assessing whether covariates followed the proportional hazards assumption, the Schoenfeld residual analyses and log-log plots suggested that some of the variables violated this assumption. However, since this research uses a reasonably large dataset there is a lot of power to detect small deviations. To deal with the potential non-proportional hazards for some of the covariates, a time dependent variable can be included for the non-proportional predictors or the model can be stratified by the predictor. An alternative solution is to fit a model which does not have the proportional hazards assumption.

The recording of cancer diagnoses could be enhanced using data linkage to cancer registries. For example, cause of death from the ONS could be used to improve reliability of results. Electronic recording of the cause of death has been shown to be incomplete even when free text entries are reviewed.[81]

Not all patients will go to see their GP if they have symptoms, this study focussed on reported symptoms which could underestimate the number of people who actually have symptoms. This would mean that HRs could be underestimated if they are not reported. The screening population are different to those who present to primary care. However, steps were made during this analysis to define a screening cohort by limiting to those aged between 60-74 and those with a BCSP FOBT electronic result. The test accuracy measures,

sensitivity and specificity, of the FOBT were similar to those reported in other studies providing a layer of data validity.

Limitations specific to primary care databases are that the data are collected during routine practice or consultation and therefore not primarily collected for research. There are several limitations arising from this, particularly in relation to missing data.[82] For example, in primary care records there may be recording errors, incomplete data, misclassification (e.g. ex smoker as non smoker),[83] variations in GP practice recording (Read codes, different operating systems, administrative procedures) or missed information from secondary care and pharmacies. Furthermore, data recording practices can change from outside influences such as the Quality Outcomes Framework (QOF) or from NICE guidance in England. Pathology and other laboratory test results (e.g. bowel cancer screening test results) were previously sent by letter, this has evolved over time due to increasing computerisation into electronic notifications, which may be considered more complete and less biased towards the recording of positive results. Furthermore, the frequency of the 'Anaemia' Read code in a recent prediction modelling study was found to decline in use over time, with a corresponding increase in the recording of a haemoglobin result.[84] GP behaviour will also affect coding of information, for example they may have a preference to use certain codes or to record certain prescriptions or to justify the prescription of a particular drug. During a consultation, a GP may not record every symptom but may record the key features associated with a diagnosis. 'Change in Bowel habit', a predictor for colorectal cancer for instance was found to be used differently to recording a symptom of diarrhoea or constipation by GPs.[84]

Incomplete data can lead to bias in parameter estimates or in sample selection (selection bias). Confounder variables and health indicators often have missing data.[80] There are several strategies to deal with missing data which may give more precise parameter estimates and reduce selection bias.[85] Multiple imputation for instance accounts for the uncertainty of missing data by producing multiple imputed datasets and combining the results across these datasets.[85] [86] This statistical approach can be used if the data are considered 'missing at random' (MAR) and leads to more accurate standard errors and p-values compared to other methods.[87] A 'compete cases' approach is often used by researchers, which can lead to biased estimates of predictor-outcome associations if missingness is associated with the outcome. Biased model performance estimates when compared to using the whole dataset and can also affect generalizability of the model.[87]

The removal of subjects due to the missingness of several different variables can lead to a reduction in the sample size causing a loss of power and therefore less precise parameter estimates. Other statistical approaches for handling missing data include creating a 'missingness category' for the variables with missing data. This approach can fail to adjust correctly for confounding, introducing bias and is generally not recommended.[87 88] Creating a 'missingness category' also leads to the categorisation of continuous predictors which in turn leads to a loss of information and is not recommended for risk prediction model development.[89 90] Other strategies which are also considered biased include 'last observation carried forward', or 'imputing the mean' determined from the observed data; these approaches however do not account for the uncertainty of the missing values leading to standard errors which are underestimated.[85]

Although multiple imputation is considered the least biased approach to handle missing data, there are issues with applying this method when using primary care data because the information is recorded for a clinical reason. For instance, a study investigating missing data in THIN and comparing results obtained using multiple imputation to two nationally representative datasets (The Health Survey for England and The British Regional Heart Survey) found that data was not missing at random for smoking and alcohol consumption.[80] A further example as described above is the use of lab test results or pathology data before the introduction of electronic notifications. Positive results from these laboratory tests were more likely to be recorded than those with normal results which would bias predictor-outcome associations. When data is not missing at random, it is suggested that sensitivity analyses investigating different missingness assumptions could be conducted; this approach could be considered in future EHR research.[85] Missing data present in this study was reported fully using study flow diagrams and includes percentage completeness of variables so researchers can assess the applicability and generalizability of study findings for future prediction models.

4.3 Strengths and weaknesses in relation to other studies

The only study identified as using survival analysis modelling techniques from the systematic review in **Chapter 2** was by Yen *et al.*[1] . For this study, the FIT was combined with conventional risk factors obtained from a questionnaire along with lab results (triglyceride levels). This study used a cohort from Taiwan invited to population-based screening for colorectal neoplasia (n=54,921) to develop the model and another two datasets combined for external geographical validation (n=17,085). For the model

combining FIT with other factors to predict colorectal neoplasia, the AUC ROC was 0.835 (95% CI: 0.821-0.849) in the development dataset and 0.861 (0.852-0.869) in the validation dataset. Although not directly comparable as this used an accelerated failure time model, the C statistic for the present study was 0.850 (adjusted for optimism) for the model which combined the FOBT with other risk predictors. The gFOBT has lower test accuracy than FIT but still had similar discriminatory power when including the other routinely available predictors. This study also has the advantage of using routine data instead of requiring further lab tests and questionnaires.

There is differential verification of cancer in this dataset because it is real world data. Patients with a positive FOBT will have been offered further testing, increasing the probability of detecting cancer if present. Patients with a negative FOBT are less likely to have cancer if present, because they are likely to have fewer follow up tests over the 2 year period. Therefore, the model may overestimate the predictive power of FOBT and other variables used in the current pathway to determine whether to refer for colonoscopy, and underestimate the predictive power of those variables not used in the referral pathway. This is a necessary limitation of using routine data. Therefore this model can be used to highlight potential predictors for future models, but not necessarily to calculate absolute predictive ability due to the effects of the underlying screening pathway on model parameters.

A survival analysis approach was taken to enable use of the longitudinal data, rather than a logistic regression model based on the sum of events in the 2 year period. Other similar studies which have investigated the risk of colorectal cancer in primary care from the presence of clinical features have used logistic regression to quantify the associations.[20 21] Survival analysis is a more efficient use of longitudinal primary care data which allows us to maximise events at the tail end and can take individual follow up into account compared to logistic regression. In addition, cancer detected soon after testing is more relevant to screening than a cancer detected 2 years later. On the other hand, the disadvantage of this approach is the greater weight given to earlier events as a result of follow up tests post positive FOBT, so may increase the effects of differential verification.

Yen *et al*.[1] adjusted the time to event for individuals with a FIT result of ≥100 ng/ml due to the earlier detection of colorectal cancer/adenoma within these individuals. A correction factor was applied to adjust the time to event for these individuals, a similar approach could be considered for the model developed in this study as those with positive FOBTs

would have quicker diagnosis due to the screening pathway. The dichotomised outcome makes this more complex but could be considered in future iterations.

Perhaps the most comparable study is the model developed by Hippisley-Cox and Coupland[15] who derived a risk prediction model using the QResearch database and Cox Regression to predict current colorectal cancer. This model was developed for use in a primary care setting to facilitate early referral for patients at high risk of existing colorectal cancer. Predictor variables retained in the final model included, age, family history of gastrointestinal cancer, anaemia, rectal bleeding, abdominal pain, appetite loss and weight loss (alcohol status and recent change in bowel habit were also significant for males). These predictors retained in the final model are similar to the predictors included in the final model in the current study. The AUC ROC was 0.89 for females and 0.91 in males in the validation sets. This has slightly higher performance than the current study (C statistic: 0.850) but this model is used in a screening setting to predict both colorectal cancer and polyps. Both the current study and the study by Hippisley-Cox and Coupland[15] estimate the baseline hazard by setting the covariates equal to zero, it may be better to centre the predictors by their mean in order to get the survival probability with the average of these predictors/characteristics for a person with average characteristics (e.g. age, gender, symptoms).

4.4 Practical implications

This is an exploratory analysis investigating the potential use of GP records in informing screening based decisions for Colorectal Cancer. The FIT will be implemented in Summer 2018 in the NHS Bowel Cancer Screening Programme and this test will most likely have its own set of SNOMED codes to electronically notify GP practices about results. After a few years of follow up of this test, the methods used in this analysis can be repeated to look at the potential impact of combining the FIT (continuous test) with other risk factors (as opposed to the FOBT) which may enhance model performance and test accuracy further.

Most factors retained in the multivariable model have a high level of recording. Laboratory parameters are the factors least well recorded and would only be requested if a GP suspects underlying abnormalities based on symptoms. Evidence suggests that blood results are important indicators of underlying cancers and other conditions.[24-27] Other studies have shown the merit of using blood test results combined with screening tests.[28 30] FOBTs on their own may miss intermittent or low level bleeding whereas a blood

parameter such as anaemia (Hb Concentration) may detect these scenarios. The added effect of lab data may help to reduce false negatives and false positives from the screening test. Routine blood test results for those in the screening age range could be implemented in the future. For example, the NHS Health Check is offered to individuals aged 40-74 and this could include routine blood tests.

After a positive FOBT, individuals are referred on to a specialist screening practitioner to discuss the positive result and suitability for colonoscopy. Not all individuals attend the SSP clinic (around 6%) and not all even if suitable for colonoscopy go on to have the diagnostic test (83% attend diagnostic examination).[77] By having an indication of an individual's risk of bowel cancer being diagnosed at colonoscopy compared to average, this could help both the patient and screening practitioner make a more informed clinical decision. Risk information can be presented to potentially help informed decision making and increase uptake of the reference standard as well as the screening test.[91 92] If an individual has a lack of perceived risk, this can affect uptake. A nomogram, such as the one presented for the multivariable model in **Section 3.6.3** could be used for this purpose.

The choice of a parametric model over a semi-parametric one in this instance depends on several aspects including model performance parameters, clinical scenario, external validation and practical application. The Cox model has the most flexibility and is the most commonly implemented model in the literature with fully developed methods and applications. Furthermore, Cox Regression could be preferred for out of sample validation due to published guidance on the appropriate methods.[63] Parametric models on the other hand can provide more precise parameter estimates provided the underlying assumptions are true and generally provide smooth estimates of the hazard and survival function.[33] They allow the baseline hazard to be determined at many different time points enhancing post-estimation measures. From a prediction modelling standpoint, this allows predictions/probabilities to be estimated at a range of different time values (e.g. 6 months, 1 year, 2 years) with potentially more accuracy.

In this study, the Nelson-Aalen cumulative hazard increases constantly for about 550 days and then has a sharper increase for the last 180 days for those with negative results. The performance of the parametric models compared to the Cox regression model was very similar and so the choice of model may come down to individual considerations. The covariate effect estimates for the scenario with negative FOBT results only showed very similar results between the Gompertz model and the Cox Regression model. The

generalized gamma model has three parameters so can fit the tail end of the data better compared to the other model types. The computational time was significantly longer when fitting the generalized gamma model over the other AFT models (and Cox regression) for a population with positive and negative FOBT results which can suggest difficulty with model convergence. External validation of model performance may show more difference between model types, although internal validation with bootstrapping did not show much difference between the models in this instance. In order to make out of sample predictions for the Cox model special measures such as interpolation or extrapolation are required which can limit its application.[33] Flexible parametric models allow the hazard to be modelled more closely and can provide more accurate parameter estimates and predictions. If the parametric models showed a significant improvement in model performance then this could have provided further justification to investigate a flexible parametric model. These models tend to have a better fit but can be difficult to interpret and may overfit to the data which will have repercussions on external validation performance and generalisability.

4.5 Future Research

When patients sign up to a GP practice, their details are uploaded to the NHS Information Authority. The NHS Spine draws out these registration details over-night. Correspondingly the details of everyone who falls within the age range of screening (60-74) is extracted to the BCSS overnight to gain new patient details. The Spine is set to what information is drawn from the GP practice, but there is capacity to draw out additional information to the BCSS. The factors shown in this study to be predictive of colorectal cancer could be considered in the future to combine with the screening test to identify those at highest risk and who would benefit most from colonoscopy.

This approach could be considered for a future risk based project once the FIT is implemented and the new SNOMED coding system defined for this test. There would be many issues relating to quality of data, how the data are handled by the GP and the use of different GP operating systems in the NHS. Health systems in the UK are also transitioning over to SNOMED CT clinical terminology and so Read codes will eventually cease to be used. As the FIT is implemented in the UK, GP records will begin to obtain more data on the FIT. The approaches used in this chapter could be used to investigate prediction models combining this newer test.

Although parametric models were investigated to determine whether these provided a better fit compared to using a Cox Regression model, the baseline hazard shape could be modelled better (particularly for the model combining FOBT with risk predictors) using a flexible parametric survival model. Royston-Parmar models are flexible parametric models which use restricted cubic splines to model the baseline hazard more closely.[33] This model can be used to derive hazard ratios which give similar results to Cox Regression and can be used to produce absolute risk predictions.

Personalised screening intervals could be investigated by plotting Kaplan Meier curves for time to diagnosis over 5 years stratified by a risk prediction model with different patterns of covariates. There is research to suggest, that the haemoglobin concentrations detected by the FIT are associated with the detection of adenomas in future screening rounds.[93] Other factors combined with the FIT could also be used to examine detection in future screening rounds.

Longitudinal recording of laboratory test results and FIT screening results may provide an additional layer of information contributing to a risk score. This study considered predictors which were available/measured at the time of entry to the study with a time limit to ensure that the predictor was associated with the outcome if recorded. Further risk information can be derived by looking at the longitudinal change in certain parameters for instance lab test results. These are time dependent covariates and there are other survival analysis methods which can be used to capture this information over time.

These models were developed for an English screening population utilising the NHS BCSP and the BCSS to record screening information. Scotland, Wales and Ireland have their own IT systems and procedures to notify GPs of results. In addition, Read codes are utilised in different ways according to the regions. Future research could look at developing models for different regions which would require different Read codes/clinical codes and IT screening systems.

5.0 CONCLUSIONS

This chapter has shown that there are several clinical predictors available from GP databases which are associated with colorectal cancer and polyps for an English average risk screening population. Furthermore, this research has identified predictors which could be considered for inclusion in a future risk adjusted screening model. Most factors contributing to the risk based models are well recorded. Laboratory parameters although

shown to be associated with colorectal cancer diagnosis are the least well recorded factors included in the final risk prediction model. Further data could potentially be drawn from primary care onto the BCSS for use in screening referral algorithms. Additional predictors which could be considered for inclusion in a future risk adjusted model include those which relate to screening history which have a strong association with the diagnosis of colorectal cancer/polyps (previous positive FOBT results and previous polyps diagnosed). Previous polyps diagnosed retained significance in both multivariable models. Interestingly, lifestyle factors, alcohol consumption and smoking status, were significant in the model for positive and negative FOBT results but just smoking in the model for negative FOBTs only (most likely explained by other variables within this model). Family history of gastrointestinal cancer was a strong predictor independently and in both multivariable models. Other variables common to both models included demographic characteristics age and sex as well as the following symptoms: abdominal pain/antispasmodic prescription, diarrhoea, change in bowel habit and flatulence. Finally the blood test result for MCV was retained in both models. Similar analyses could be carried out with the FIT which is due to be introduced to the NHS BCSP along with corresponding SNOMED codes. This chapter identifies additional predictors for consideration in a future risk based screening model and helps to decide if someone with a negative FOBT would benefit from further investigation if they are at higher risk from other predictors.

6.0 REFERENCES

1. Yen AM, Chen SL, Chiu SY, Fann JC, Wang PE, Lin SC, et al. A new insight into fecal hemoglobin concentration-dependent predictor for colorectal neoplasia. Int J Cancer. 2014;135(5):1203-12.
2. Chaudhry Z, Mannan F, Gibson-White A, Syed U, Ahmed S, Kousoulis A, et al. Outputs and Growth of Primary Care Databases in the United Kingdom: Bibliometric Analysis. J Innov Health Inform. 2017;24(3):942.
3. IMS Health. IMS Health Statistics 2017 [Available from: http://www.csdmruk.imshealth.com/our-data/statistics.shtml.
4. IMS Health. THIN Data Guide for Researchers 3.0. 2015.
5. UK National Screening Committee. Bowel cancer screening across the UK 2013 [Available from: http://www.screening.nhs.uk/bowelcancer-compare.
6. Launois R, Le Moine JG, Uzzan B, Fiestas Navarrete LI, Benamouzig R. Systematic review and bivariate/HSROC random-effect meta-analysis of immunochemical and guaiac-based fecal occult blood tests for colorectal cancer screening. European journal of gastroenterology & hepatology. 2014;26(9):978-89.
7. NHS Digital. Summary Care Records 2017 [Available from: https://digital.nhs.uk/summary-care-records.
8. NHS Digital. Spine 2017 [Available from: https://digital.nhs.uk/spine.
9. NHS Connecting for Health. Bowel Cancer Screening Programme - Electronic Results Communication. 2010.
10. Stegeman I, de Wijkerslooth TR, Stoop EM, van Leerdam ME, Dekker E, van Ballegooijen M, et al. Combining risk factors with faecal immunochemical test outcome for selecting CRC screenees for colonoscopy. Gut. 2014;63(3):466-71.
11. Auge JM, Pellise M, Escudero JM, Hernandez C, Andreu M, Grau J, et al. Risk Stratification for Advanced Colorectal Neoplasia According to Fecal Hemoglobin Concentration in a Colorectal Cancer Screening Program. Gastroenterology. 2014.
12. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). Br J Cancer. 2012;107(2):260-5.
13. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. The British journal of general practice : the journal of the Royal College of General Practitioners. 2013;63(606):e11-21.
14. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. The British journal of general practice : the journal of the Royal College of General Practitioners. 2013;63(606):e1-10.
15. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. The British journal of general practice : the journal of the Royal College of General Practitioners. 2012;62(594):e29-37.
16. Fijten GH, Starmans R, Muris JW, Schouten HJ, Blijham GH, Knottnerus JA. Predictive value of signs and symptoms for colorectal cancer in patients with rectal bleeding in general practice. Family practice. 1995;12(3):279-86.
17. Wauters H, Van Casteren V, Buntinx F. Rectal bleeding and colorectal cancer in general practice: diagnostic study. BMJ (Clinical research ed). 2000;321(7267):998-9.
18. Hamilton W, Round A, Sharp D, Peters T. Clinical features of colorectal cancer before diagnosis: a population-based case-control study. Br J Cancer. 2005;93:399 - 405.
19. Selvachandran SN, Hodder RJ, Ballal MS, Jones P, Cade D. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. The Lancet. 2002;360(9329):278-83.

20. Hamilton W. The CAPER studies: five case-control studies aimed at identifying and quantifying the risk of cancer in symptomatic primary care patients. Br J Cancer. 2009;101 Suppl 2:S80-6.
21. Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut. 2011;60(9):1242-8.
22. NICE. Referral guidelines for suspected cancer 2005 [Available from: http://www.nice.org.uk/nicemedia/live/10968/29814/29814.pdf.
23. Kidney E, Berkman L, Macherianakis A, Morton D, Dowswell G, Hamilton W, et al. Preliminary results of a feasibility study of the use of information technology for identification of suspected colorectal cancer in primary care: the CREDIBLE study. Br J Cancer. 2015;112(s1):S70-S6.
24. Goldshtein I, Neeman U, Chodick G, Shalev V. Variations in hemoglobin before colorectal cancer diagnosis. European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP). 2010;19(5):342-4.
25. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng KK, Marshall T. The importance of anaemia in diagnosing colorectal cancer: a case-control study using electronic primary care records. Br J Cancer. 2008;98(2):323-7.
26. Spell DW, Jones DV, Jr., Harper WF, David Bessman J. The value of a complete blood count in predicting cancer of the colon. Cancer detection and prevention. 2004;28(1):37-42.
27. Bailey SE, Ukoumunne OC, Shephard EA, Hamilton W. Clinical relevance of thrombocytosis in primary care: a prospective cohort study of cancer incidence using English electronic medical records and cancer registry data. The British journal of general practice : the journal of the Royal College of General Practitioners. 2017;67(659):e405-e13.
28. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. Journal of the American Medical Informatics Association : JAMIA. 2016;23(5):879-90.
29. Birks J, Bankhead C, Holt TA, Fuller A, Patnick J. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer medicine. 2017;6(10):2453-60.
30. Boursi B, Mamtani R, Hwang WT, Haynes K, Yang YX. A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results. Digestive diseases and sciences. 2016;61(7):2076-86.
31. Spell DW, Jones DV, Harper WF, David Bessman J. The value of a complete blood count in predicting cancer of the colon. Cancer Detection and Prevention. 2004;28(1):37-42.
32. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 2003;89(2):232-8.
33. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. Texas, United States: Stata Press Publication; 2011.
34. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis - an introduction to concepts and methods. Br J Cancer. 2003;89(3):431-6.
35. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ : British Medical Journal. 2007;335(7611):136-.
36. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. PLoS Med. 2007;4(10):e296.

37. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015;12(10):e1001885.
38. McManus RJ, Ryan R, Jones M, Wilson S, Hobbs FR. How representative of primary care are research active practices? Cross-sectional survey. Family practice. 2008;25(1):56-62.
39. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiol Drug Saf. 2009;18(1):76-83.
40. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. Br J Cancer. 2003;89(4):605-11.
41. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. Journal of clinical epidemiology. 1995;48(12):1495-501.
42. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of clinical epidemiology. 1995;48(12):1503-10.
43. Steyerberg EW. Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer; 2009.
44. IMS Health. Ethics 2017 [Available from: http://www.epic-uk.org/our-data/ethics.shtml.
45. National Institute for Health and Care Excellence. Suspected cancer: recognition and referral [NICE guidelines NG12]June 2015 19th January 2016. Available from: http://www.nice.org.uk/guidance/ng12.
46. Hamilton W, Lancashire R, Sharp D, Peters T, Cheng K, Marshall T. The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. BMC Medicine. 2009;7(1):17.
47. Williams TGS, Cubiella J, Griffin SJ, Walter FM, Usher-Smith JA. Risk prediction models for colorectal cancer in people with symptoms: a systematic review. BMC Gastroenterology. 2016;16:63.
48. Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiology and drug safety. 2009;18(8):704-7.
49. Excellence NIfHaC. Quantitative faecal immunochemical tests to guide referral for colorectal cancer in primary care (DG30). nice.org.uk/guidance/dg30; 2017.
50. Welch CA. Ethnicity UCL - The Health Improvement Network (THIN) Research Team Webpage2012 [Available from: https://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/mi/recording_in_thin/ethnicity.
51. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, White IR, et al. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. BMJ Open. 2014;4(4).
52. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53:457-81.
53. Moons KM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.
54. Efron B. The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association. 1977;72:557-65.
55. Stata.com. mfp — Multivariable fractional polynomial models 2017 [Available from: https://www.stata.com/manuals13/rmfp.pdf.

56. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. Computational Statistics & Data Analysis. 2006;50(12):3464-85.
57. Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology--with an emphasis on fractional polynomials. Methods Inf Med. 2005;44(4):561-71.
58. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982;247(18):2543-6.
59. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine. 1996;15(4):361-87.
60. Cleves M, Gould W, Marchenko YV. An Introduction to Survival Analysis Using Stata. Revised third edition ed. Texas, USA: Stata Press; 2016.
61. Korn EL, Simon R. Measures of explained variation for survival data. Statistics in medicine. 1990;9(5):487-503.
62. Cox D. Note on Grouping. J Am Stat Assoc. 1957;52:543-7.
63. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
64. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Med. 2010;8:21.
65. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008;167(3):362-8.
66. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Medicine. 2013;10(2):e1001381.
67. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ (Clinical research ed). 2016;353.
68. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Statistics in medicine. 1990;9(11):1303-25.
69. Akaike H. Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, ed B N Petrov and Csaki Budapest: Akademiai Kiado. 1973:267-81.
70. Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978;6(2):461-4.
71. Zlotnik A, Abraira V. A general-purpose nomogram generator for predictive logistic regression models. Stata Journal. 2015;15(2).
72. Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. The Stata Journal. 2010;10(3):339-58.
73. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Statistics in medicine. 2004;23(5):723-48.
74. Royston P. Explained Variation for Survival Models. Stata Journal, StataCorp LP. 2006;6(1):83-96.
75. StataCorp. streg - Parametric survival models College Station, TX: Stata Press2017 [Available from: https://www.stata.com/manuals13/ststreg.pdf.
76. Rees CJ, Bevan R. The National Health Service Bowel Cancer Screening Program: the early years. Expert review of gastroenterology & hepatology. 2013;7(5):421-37.
77. Logan RFA, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. Gut. 2012;61(10):1439-46.

78. Royston P. STR2D: Stata module to compute explained variation for survival models. Statistical Software Components. 2011;S457228, Boston College Department of Economics.
79. Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society, series B. 1968;30(2):248-75.
80. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiology and drug safety. 2010;19(6):618-26.
81. Hall GC. Validation of death and suicide recording on the THIN UK primary care database. Pharmacoepidemiology and drug safety. 2009;18(2):120-31.
82. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annual review of public health. 2016;37:61-81.
83. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, White IR, et al. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: a cross-sectional study. BMJ Open. 2014;4(4):e004958.
84. Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut. 2011;60(9):1242.
85. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clinical research ed). 2009;338.
86. Carpenter JR, Kenward MG. Multiple Imputation and Its Application (Statistics in Practice). 1st edition ed. Chichester, West Sussex: John Wiley & Sons; 2013.
87. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
88. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006;59(10):1087-91.
89. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.
90. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ (Clinical research ed). 2006;332(7549):1080.
91. van Vugt HA, Roobol MJ, Venderbos LD, Joosten-van Zwanenburg E, Essink-Bot ML, Steyerberg EW, et al. Informed decision making on PSA testing for the detection of prostate cancer: an evaluation of a leaflet with risk indicator. European journal of cancer (Oxford, England : 1990). 2010;46(3):669-77.
92. Edwards AG, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. The Cochrane database of systematic reviews. 2013(2):Cd001865.
93. Digby J, Fraser CG, Carey FA, Diament RH, Balsitis M, Steele RJ. Faecal haemoglobin concentration is related to detection of advanced colorectal neoplasia in the next screening round. J Med Screen. 2017;24(2):62-8.

7.0 APPENDICES

Appendix 1: Frequency of Read codes used to diagnose bowel cancer from the THIN database.

Read code Description	Frequency of Code used to record Bowel Cancer
Bowel cancer detected by national screen	20
Bowel scope (flexible sigmoidoscopy) scr	1
Cancer of bowel	112
Carcinoma of caecum	4
Carcinoma of rectum	15
Colonic cancer	83
H/O Lower GIT Neoplasm	2
Malig neop other site rectum, rectosigmo	2
Malignant neoplasm of appendix	8
Malignant neoplasm of ascending colon	38
Malignant neoplasm of caecum	59
Malignant neoplasm of colon	314
Malignant neoplasm of colon NOS	56
Malignant neoplasm of descending colon	11
Malignant neoplasm of hepatic flexure of	8
Malignant neoplasm of other specified si	2
Malignant neoplasm of rectosigmoid junct	10
Malignant neoplasm of rectum	202
Malignant neoplasm of rectum, rectosigmo	15
Malignant neoplasm of retrocaecal tissue	1
Malignant neoplasm of sigmoid colon	120
Malignant neoplasm of splenic flexure of	8
Malignant neoplasm of transverse colon	11
Malignant neoplasm rectum,rectosigmoid j	1
Rectal carcinoma	102
[M]Adenocarcinoma in adenomatous polyp	1
[M]Adenocarcinoma in multiple adenomatou	1
[M]Adenocarcinoma in tubulovillous adeno	3
[M]Pseudomyxoma peritonei	7
[M]Tubular adenocarcinoma	1
[M]Tubular adenoma or adenocarcinoma NOS	1
[M]Tubular adenomas and adenocarcinomas	68
[M]Villous adenomas and adenocarcinomas	5

Table A.1.1: The frequency of Read codes used to diagnose colorectal cancer from the THIN database. The Bowel Cancer Screening Programme Cancer Diagnosis code cannot be relied upon to distinguish between screen detected and clinically detected cancers based on the frequencies listed in the table. This is based on a 1% sample from the THIN database.

Appendix 2: Eligibility Criteria for Data Extraction

Practice Criteria

Practice Criteria	Notes
Practice Start Date: The latest of AMR, ABSD (Acceptable bowel cancer screening date – defined by researcher) and Vision date plus one year.	ABSD date is saved as: practice_screening_start_date.dta
Practice End Date: The last collection date from each practice	
Exclude practices not in England	
Exclude practices where the practice start date is greater than or equal to the practice end date	Also dealt with later in the exclusion process when patients whose start date is >= their end date are excluded

Table A.2.1: Practice eligibility criteria for data extraction.

Patient Criteria

Patient Criteria	Notes
Patient Observation Start: Later of Practice start date, registration date plus one year and age 59.	Operationalised during data extraction as the beginning of the year in which the patient turns 60. This includes some period of time when they were aged 59.
Patient Observation End: Earlier of Practice end date, De-registration or death and age 75.	Operationalised during data extraction as the beginning of the year in which the patient turns 75. This includes some period of time when they were aged 75.
Period Start (Patient Index Date) Latest Bowel Cancer Screening Programme FOBT result	AHD results only lk_BCSP_FOBT_codes.dta
Period End CRC/Polyp (Patient End Date): Earlier of patient observation end and outcome of interest (colorectal cancer diagnosis)	lk_colorectal_cancer.dta lk_polyp.dta Generate from data
Exclude if not permanently registered or applied for permanent registration	(patflag)
Exclude if patient observation start is greater than or equal to observation end	
Exclude if sex is not equal to male or female.	
Exclude if no screening FOBT record	
Exclude if CRC diagnosis prior to period start	lk_colorectal_cancer.dta
Exclude patients at high risk of bowel cancer due to inherited conditions - FAP (familial adenomatous polyposis) and Lynch syndrome (hereditary non polyposis colorectal cancer or HNPCC) – Read coded prior to FOBT.	lk_HNPCC.dta lk_familial_adenomatous_polyposis.dta

Table A.2.2: Patient eligibility criteria for data extraction.

Appendix 3: Scientific Review Committee Approval Letter

SRC Feedback

Researcher Name: Jennifer Cooper
Organisation: Warwick Medical School
SRC Reference Number: 16THIN037
Date: 26/05/2016
Study title: The feasibility and accuracy of using routinely recorded data from electronic GP records in a risk prediction model combining the FOBT for colorectal cancer screening

Committee opinion: Approved

The following feedback has been supplied by the SRC.

Notes from the Chair:

Advice (General advice for the researchers as information only)
A) Please be aware that the study power could be very limited for some sub-group analyses propsed in this study. Sample size calcuations for sub-group analyses would be helpful in this case
B) If the aim is to refine a risk prediction model for bowel cancer using the screening population (60-74 years) should the age cut off for inclusion into the study be more rigid ie.not include those aged 30 years. Although the numbers in the 30-60 age group having FOBT performed may be small it is likely that these will be a different population (those with familial risk factors such as polyposis coli or IBD).

We are pleased to inform that you can proceed with the study as this is now approved. IMS Health will let the relevant Ethics committee know this study has been approved by the SRC.

Once the study has been completed and published, it is important for you to inform IMS Health in order for us to advise the SRC and your reference number to be closed.

References to all published studies are added to our website enabling other researchers to become aware of your work. In order to identify your study as using the THIN database, we recommend that you include the words "The Health Improvement Network (THIN)" within your title. Copies of publication(s), where available, will be appreciated.

Studies using THIN-HES linked data are required to include the following statement as per agreements with HSCIC:

"Copyright © 2015, Re-used with the permission of The Health and Social Care Information Centre. All rights reserved."

I wish you and your team all the best with the study progression.

Mustafa Dungarwalla
Consultant

Page 1

SRC
Scientific Review Committee

Appendix 4: Variable Definitions/Specification for Data Extraction

Variable Name	Definition
BMI	Latest record prior to FOBT
Weight	Latest record prior to FOBT Also can generate weight loss - from the most recent and previous weights. Calculated difference between two weights (<5% drop, 5 to 9.9% drop or >=10% drop)
Height	Latest record prior to FOBT
Smoking status	Latest record prior to FOBT
Alcohol	Latest record prior to FOBT Units per week. Alcohol status can be worked out from AHD-recorded number of units per week, AHD Read code associated with the AHD code for alcohol consumption and a separate MEDICAL table Read code lookup used by QRisk2: 'lk_read_alcohol.dta' Alcohol status (non-drinker; trivial [<1 unit/day]; light [1–2 units/day]; moderate/heavy [≥3 units/day])
Mean cell volume	Continuous measure defined in AHD document All records prior to FOBT For analysis the latest record within 365 days so it is associated with the outcome.
Ferritin	Continuous measure defined in AHD document All records prior to FOBT For analysis the latest record within 365 days so it is associated with the outcome.
Blood group	Defined in AHD document
Family history of colorectal cancer	Defined in AHD document
Platelet count	Continuous measure defined in AHD document All records prior to FOBT For analysis the latest record within 365 days so it is associated with the outcome.
FOBT Screening Outcome	Defined in AHD document
Primary FOBT result	Defined in AHD document
Haemoglobin concentration	Continuous measure defined in AHD document All records prior to FOBT For analysis the latest record within 365 days so it is associated with the outcome.

Table A.4.1: AHD variable definitions for data extraction.

Variable Name	Definition
Abdominal mass	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Abdominal pain	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Abnormal rectal examination	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.date.
Change in bowel habit	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Colorectal Cancer Diagnosis	Date of the first after the FOBT
Constipation	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Crohn's Disease	Date of first diagnosis if prior to the FOBT date
Diabetes	Up to 365 days prior to the index date Date of first diagnosis if prior to the FOBT date Standard list of diabetes codes. Both types
Diarrhoea	For data extraction: Date of the most recent prior to the FOBT date.

	Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Diverticulitis/Diverticulosis	Only bowel diverticula For data extraction: Date of the most recent prior to the FOBT date if ever recorded.
Family History of Colorectal Cancer	If ever recorded prior to FOBT date
FAP (familial adenomatous polyposis)	As an exclusion factor (these patients are at higher risk than the average risk screening population) First ever record
Ferritin (low levels)	Also investigated in AHD table Latest recorded prior to FOBT
Flatulence	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Flexible Sigmoidoscopy Record (BCSP)	These codes should identify patients who have a record from the bowel cancer screening programme relating to flexible simoidoscopy
FOBT Result (Primary Care)	Also investigated in the AHD table Up to 365 days before latest FOBT result.
FOBT Screening Outcome (Latest FOBT and previous FOBT tests)	These codes should identify patients who have a record from the bowel cancer screening programme relating to the FOBT - allowing an individual's screening history to be extracted. Also investigated in AHD table
Hereditary Nonpolyposis Colorectal Cancer (HNPCC)	As an exclusion factor (these patients are at higher risk than the average risk screening population) First ever record
IBS (Irritable Bowel Syndrome)	For data extraction: Date of the most recent prior to the FOBT date if ever recorded.
Loss of appetite	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Obesity	Patients who have had an obesity diagnosis within 2 years of the index date. Also investigated in AHD table
Polyp diagnosis (Prior to FOBT and after)	Date of the most recent prior to the FOBT date if ever recorded. Date of the first after the FOBT date.
Rectal bleeding/Melaena	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Thrombocytosis	Platelet count will also be extracted from the AHD table For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Tiredness	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Ulcerative Colitis	Date of first diagnosis if prior to the FOBT date if ever recorded
Venous Thromboembolism (VTE) which includes Pulmonary Embolism and Deep Vein Thrombosis	For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.
Weight loss	Also investigated in AHD table For data extraction: Date of the most recent prior to the FOBT date. Up to 365 days prior to the index date: Generate binary variable to indicate if the exposure is recent.

Table A.4.2: Read code lookup variable definitions for data extraction.

Variable Name	Definition
Anti-motility drugs (proxy for diarrhoea)	Drug listed in drug_lk_antimotility Up to 365 days before latest FOBT
Laxatives (proxy for constipation)	Drug listed in drug_lk_laxative Up to 365 days before latest FOBT
Antispasmodics (proxy for abdominal pain)	Drug listed in drug_lk_antispasmodic Up to 365 days before latest FOBT

Table A.4.3: Drug Code lookup variable definitions for data extraction.

Patient File (Variable Name)
Patient ID
Practice ID
Registration date
AMR date
Vision date
Date of death
Date left practice (de-registration)
Year of birth
Sex
Practice File (Variable Name)
AMR date
Vision date
Last collection date
PVI File Lookups (Variable Name)
Townsend Quintile
Ethnicity

Table A.4.4: Data required from additional files in THIN: Patient file, Practice file, PVI file.

Appendix 5: Table contents for Data Extraction and Analysis

Variable	Notes
Patient ID	Patient file
Practice ID	Patient file
Registration date	Patient file
AMR date	Practice file Included in the definition of the variable start_date
Vision date	Practice file Included in the definition of the variable start_date
Screening start date (Acceptable bowel cancer screening date)	practice_screening_start_date.dta
Last collection date	Practice file Included in the definition of the variable end_date
Date of death	Patient file
Date left practice (de-registration)	Patient file
Year of birth	Patient file
Sex	Patient file
Ethnicity	PVI
Family history of Colorectal Cancer	Latest record prior to FOBT AHD Medical Table
Height	Latest record prior to FOBT AHD
Weight	Latest record prior to FOBT AHD
BMI	Latest record prior to FOBT AHD Two columns for this: bmi_vision – The BMI recorded in the patient record bmi_calc – Recalculated BMI based on the weight recorded on that day and the median recorded height for that patient.
Townsend Quintile	Latest record prior to FOBT PVI
Smoking	AHD Date Latest record prior to FOBT Status Quantity (for those that have it)
Alcohol	AHD Date Latest record prior to FOBT Units per week Latest value for units per week given in column called ahd_alcohol_units. Latest medcode in AHD table also given: ahd_alcohol_medcode

FOBT Screening outcome	AHD Medical Table (Index date) Date of the event Each Read code and data4 value (possible multiple records per patient)
FOBT Primary care outcome	AHD Medical Table Latest record prior to FOBT Each Read code and data4 value (possible multiple records per patient)
Flexible Sigmoidoscopy Record (BCSP)	Medical Table All records prior to FOBT Date (These codes should identify patients who have a record from the bowel cancer screening programme relating to flexible sigmoidoscopy screening)
Colorectal Cancer diagnosis	Medical Table Date of the first diagnosis after the screening FOBT
Polyp diagnosis prior to the FOBT	Medical Table Date of the most recent prior to the FOBT date.
Polyp diagnosis after the FOBT	Medical Table Date of the first after the FOBT date.
Blood Group	AHD Latest record prior to FOBT Date Value
Rectal bleeding/melaena	Medical Table Date of the most recent prior to the FOBT date. (NICE)
Abdominal pain	Medical Table Date of the most recent prior to the FOBT date. (NICE)
Abdominal Pain (Drug table) Antispasmodic Prescription	Drug table Most recent prior to the FOBT date
Ferritin	Medical Table (low ferritin) AHD Latest recorded prior to FOBT Date Value Additional table all ferritin dates and values. (NICE – Iron deficiency anaemia)
Mean Cell Volume	AHD Latest recorded prior to FOBT Date Value Additional table all mean cell volume dates and values.
Constipation (Read coded)	Medical Table Date of the most recent prior to the FOBT date. (NICE)
Constipation (Drug coded)	Drug table (laxative drugs) Date of the most recent prior to the FOBT date.
Diarrhoea (Read coded)	Medical Table Date of the most recent prior to the FOBT date. (NICE)
Diarrhoea (Drug coded)	Drug table (anti-motility drugs) Date of the most recent prior to the FOBT date.
Change in Bowel Habit	Medical Table Date of the most recent prior to the FOBT date. (NICE)
Weight loss	Medical Table AHD table calculation Two most recent at any time up until the FOBT date. Separate table with all the weights and corresponding dates (NICE) *Calculation - Weight loss from the most recent and previous weights. Calculated difference between two weights (<5% drop, 5 to 9.9% drop or >=10% drop)*
Rectal bleeding or Melaena (one Read code list)	Medical Table Date of the most recent prior to the FOBT date.

Crohns Disease	Medical Table Date of first diagnosis if prior to the FOBT date
Ulcerative colitis	Medical Table Date of first diagnosis if prior to the FOBT date
Haemoglobin Concentration	Latest recorded prior to FOBT Date Value Additional table all haemoglobin concentration dates and values. AHD
Diabetes	Medical Table Date of first diagnosis if prior to the FOBT date
Bowel Diverticula	Medical Table Date of the most recent prior to the FOBT date.
IBS (Read coded)	Medical Table Most recent prior to the FOBT date
Abdominal mass	Medical Table Date of the most recent prior to the FOBT date.
Abnormal rectal examination	Medical Table Date of the most recent prior to the FOBT date.
Flatulence	Medical Table Date of the most recent prior to the FOBT date.
Tiredness	Medical Table Date of the most recent prior to the FOBT date.
Venous Thromboembolism (VTE) which includes Pulmonary Embolism and Deep Vein Thrombosis	Medical Table Date of the most recent prior to the FOBT date.
Thrombocytosis	Medical Table Date of the most recent prior to the FOBT date.
Platelet Count	Latest recorded prior to FOBT Date Value Additional table all platelet count dates and values. AHD
Loss of appetite	Medical Table Date of the most recent prior to the FOBT date.
Obesity	Medical table Patients who have had an obesity diagnosis within 2 years of the index date Date of the most recent prior to the FOBT date. (Also investigated in AHD table)

Table A.5.1: Table contents for extraction

Intermediate Tables Containing:

- Ferritin
- Weight Loss
- Mean Cell Volume
- Haemoglobin Concentration
- Platelet Count
- FOBT Screening Outcome (may be multiple records per patient)
- FOBT Primary Care Outcome (may be multiple records per patient)

Patients will be followed-up until the earliest of the following dates: death; de-registration; CRC diagnosis; two years after their index date; and the date of the last data collection from their general practice.

Appendix 6: Hazard Ratios for the Cox Regression Model for a population with positive and negative FOBTs

Variable	Estimated Hazard Ratio	Bootstrapped Standard Error	z	P>z	[95% Confidence Intervals]	
MCV*age at FOBT interaction	1.090	0.046	2.04	0.04	1.004	1.184
FOBT result*age at FOBT interaction	0.964	0.014	-2.63	0.01	0.937	0.991
FOBT Result (positive result)	42.122	2.557	61.61	0.00	37.397	47.445
Smoking Status:						
ex-smoker	1.229	0.074	3.44	0.00	1.093	1.383
current smoker	1.382	0.146	3.07	0.00	1.124	1.698
Crohn's Disease Diagnosis Recorded	0.486	0.206	-1.70	0.09	0.211	1.116
Previous Polyps Diagnosed	1.912	0.269	4.60	0.00	1.450	2.520
Flatulence Symptom Recorded	2.340	1.014	1.96	0.05	1.000	5.472
MCV <80fL	1.411	0.284	1.71	0.09	0.951	2.092
Alcohol consumption (units per week)	1.086	0.033	2.67	0.01	1.022	1.153
Family History of Gastrointestinal Cancer	2.151	0.390	4.23	0.00	1.508	3.068
Abdominal pain/antispasmodic prescription recorded	1.220	0.120	2.02	0.04	1.006	1.480
Diarrhoea symptom	1.312	0.203	1.76	0.08	0.969	1.777
Sex	0.822	0.059	-2.72	0.01	0.714	0.947
Age at FOBT	1.033	0.009	3.85	0.00	1.016	1.051
Change in bowel habit symptom	2.479	0.495	4.55	0.00	1.676	3.668

Table A.6.1: Hazard ratios for final model derived from the sample population with positive and negative FOBT results.

Appendix 7: Cox Regression Diagnostics Schoenfeld Residuals

Variable	Rho	Chi^2	Degrees of freedom	Prob>chi^2
MCV*age at FOBT interaction	0.000	0.00	1	0.999
FOBT result*age at FOBT interaction	-0.013	0.19	1	0.664
FOBT Result	-0.650	536.43	1	0.000
Smoking Status:	-	-	1	-
ex-smoker	0.067	5.52	1	0.019
current smoker	0.064	5.01	1	0.025
Crohns Disease Diagnosis Recorded	0.069	5.79	1	0.016
Previous Polyps Diagnosed	0.113	15.94	1	0.000
Flatulence Symptom Recorded	0.033	1.36	1	0.243
MCV <80fL	0.019	0.44	1	0.509
Alcohol consumption (units per week)	-0.020	0.51	1	0.476
Family History of Gastrointestinal Cancer	-0.040	1.97	1	0.161
Abdominal pain/antispasmodic prescription recorded	0.005	0.03	1	0.869
Diarrhoea symptom	0.023	0.64	1	0.422
Sex	-0.019	0.48	1	0.490
Age at FOBT	0.068	5.16	1	0.023
Change in bowel habit symptom	-0.018	0.38	1	0.540
Global test		584.44	16	0.000

Table A.7.1: Test of proportional hazards using Schoenfeld Residuals. Significant results had a p-value of less than 0.05 and included; age at FOBT, previous polyps diagnosed, Crohn's disease, current smoker, ex-smoker and FOBT result. For model with positive and negative FOBTs.

Variable	rho	chi2	df	Prob>chi2
Smoking Status:	-	-	1	-
ex-smoker	0.081	3.86	1	0.049
current smoker	0.001	0	1	0.975
IBS	-0.016	0.15	1	0.695
Previous Polyps Diagnosed	0.017	0.18	1	0.676
Flatulence Symptom Recorded	0.042	1.03	1	0.310
Weight loss	-0.035	0.72	1	0.395
MCV <80fL	-0.064	2.39	1	0.122
Family History of Gastrointestinal Cancer	-0.045	1.19	1	0.276
Abdominal pain/antispasmodic prescription recorded	-0.073	3.14	1	0.076
Diarrhoea symptom	-0.072	3.02	1	0.083
Sex	0.015	0.14	1	0.708
Age at FOBT	0.109	6.54	1	0.011
Change in bowel habit symptom	-0.039	0.9	1	0.342
Global test		26.28	13	0.016

Table A.7.2: Test of proportional hazards using Schoenfeld Residuals. Significant results had a p-value of less than 0.05 and included; age at FOBT, smoking status (Ex-Smoker). For a model with negative FOBT only.

THIN Data Extraction Methodology

This research was carried out as part of an NIHR Infrastructure Doctoral Training Exchange (IDTE) Award based at the Institute of Applied Health Research at The University of Birmingham. IDTE supervisors were Professor Tom Marshall (TM) and Dr Ronan Ryan (RR) based in the Health Informatics team, Primary Care Division.

ABSTRACT

Background: The Health Improvement Network (THIN) is an anonymised GP record database derived from GP systems which use Vision software and provided for research by IMS Health. THIN provides data for over 587 practices (>5% coverage of the UK) covering more than 12 million patients. These data are arranged across four main file types linked by patient ID: patient file (for demographic information), the medical file (recording of symptoms and diagnoses), therapy file (for prescriptions) and Additional Health Data (AHD) file (laboratory test results). Diagnoses and symptoms are recorded using hierarchical Read codes and drug codes for prescriptions can be linked to the British National Formulary (BNF) Chapters. Most GP practices have electronic links with the bowel cancer screening system and so FOBT results and whether someone has participated in the programme is available in the database. Lab test results and FOBT results use the Pathlinks notification system which means test results are automatically sent through to primary care. In order to extract data from THIN for symptoms and diagnoses, Read code lists specifying the defined diagnosis/symptom need to be constructed. Similarly for prescriptions, drug code lists need to be defined. The AHD file is more complex and a strategy needs to be constructed in order to extract the data of interest, in the units of interest and within the acceptable/relevant range, particularly for laboratory test results. Finally, before approximately 2010, BCSP notifications were sent via letter to patients and practices. After this date, certain practices opted in to receive electronic notifications and a set of Read codes for this purpose was developed. In order to ensure data quality assurance for studies investigating screening programme results, dates need to be defined for individual practices to ensure electronic results are used. Before this date, practices may not have routinely recorded the results and this could have been biased towards positive results.

This Chapter aims to describe the methodology used to: define acceptable periods of BCSP notifications for practices receiving electronic results – the acceptable electronic BCSP (AEB) date, extract laboratory/other variables from the AHD file, and compile Read code

lists for diagnoses and symptoms and drug code lists for prescriptions from THIN. These methods will improve the quality and validity of the data extracted for analysis and aid transparency and reproducibility of the methods for further electronic health record (EHR) research.

Methods: For defining the AEB date, a rate of the frequency of bowel cancer screening notifications received per month by the number of patients registered in a practice aged 60-74 was determined for each practice in THIN. An expected rate for each practice was also generated based on a 50% uptake rate and people being invited biennially. Line graphs with the expected monthly rate of electronic notifications and the actual monthly rate by practice were produced along with a locally weighted scatterplot smoothing (LOWESS) line and visually examined. Rules for visual interpretation of the start date of when electronic notifications were received were compiled by 3 reviewers based on a sample of the resultant graphs. The first reviewer made a decision blind to the 2nd reviewer on when the start date would be for each practice. A consensus meeting was used to discuss the results and a third reviewer intervened where a decision could not be made. The AHD file is more complex and requires a strategy to extract valid data. The method to extract FOBT screening notifications was an iterative process which systematically reviewed the different combinations of Read codes and value labels in the AHD file observed in a THIN 1% data sample, initially restricting by the ahdcode of interest (FOBT). The AHD file was then searched for any additional BCSP codes (from the above list) not restricting by ahdcode and the output was a series of rules which can be used to identify BCSP FOBT screening outcomes for all patients in the THIN database. A similar approach was taken for defining an AHD strategy for haemoglobin concentration. The method sequentially reviewed the numeric distribution of Hb values for combinations of Read codes and value labels observed in a THIN 1% sample. These were compared with a reference distribution for Hb to assess if they matched that distribution, required conversion before use, or were unlikely to contain Hb values and therefore excluded. Plausible Hb minimum and maximum values from an external source were then applied to the data as a final step in the method. The method for Read code list generation involved an iterative key word search strategy (informed by previous research and discussion with a clinician) and then identifying other parent and child stems for Read codes by exploiting the hierarchical nature of the codes. After a few cycles of running through this, the resulting Read code list was subject to two individual reviewers (the second being a clinician for validity). The results were presented for Bowel Cancer Diagnosis. Finally, drug code list generation involved a similar method to

the above. The British National Formulary was used to identify Chapters the drugs were mapped to as well as formulate a key word search under generic drug name. All drugs mapped to the appropriate Chapter were included in the final list after removing any drugs which did not fit the definition used for the study (e.g. formulations type). The list was checked by two reviewers.

Results: The initial review of the AEB date defined for each practice gave 353 practices for inclusion and 102 for exclusion. The second reviewer gave 363 for inclusion and 92 practices for exclusion, giving 97.8% (445/455) agreement. The consensus meeting with the 3rd reviewer gave the final 92 practices for exclusion and 363 for inclusion. Eighty per cent of practices were investigated for an AEB date. The agreed AEB start dates for these practices were used in **Chapter 5** to define the population and data used for analysis. AHD strategies for bowel cancer screening notifications identified a series of rules which could be used to extract BCSP FOBT screening outcomes for all patients in the THIN database. This involved using medcodes for BCSP notifications with a definitive result. Other generic BCSP codes however needed to be combined with another ahdcode for a definitive outcome/result. The AHD strategy for haemoglobin concentration identified a reference distribution using pathology lab data and transformed values where appropriate so that the results used the same units and were recorded in the same way for analysis. Results which were outside the range of the reference distribution were excluded from extraction. Read code list generation resulted in a list of 42 codes used for bowel cancer diagnosis after being subject to a double reviewing process. Drug code list generation for laxatives resulted in a list of 450 codes after being subjected to a double reviewing process.

Conclusions:

The AEB date can be used in future studies as a layer of data quality assurance to ensure results used for analysis are ones which have been electronically received to the additional health data records. The methods derived to extract additional health data variables such as laboratory test results can be used for future studies requiring the same variables if tailoring the methods to the research question. By using an external dataset to define acceptable minimum and maximum values this also adds more validity to the data when applied in risk prediction models. There is a growing requirement to ensure Read code/Drug code lists as well as more recently their methods are transparent and available for future research studies. Data repositories exist for this purpose such as Clinical Codes set up by the University of Manchester. The strategies are fully documented in this chapter

along with code list examples at the end of this chapter. A systematic approach ensures a valid set of codes is produced for electronic health record analysis. This chapter has described the methods developed to extract data for the study in **Chapter 5** but also which can be applied in other future studies to improve data validity and quality assurance.

1.0 INTRODUCTION

The previous chapter describes a risk prediction modelling study using The Health Improvement Network (THIN) database of anonymised GP records. The objective of this chapter is to describe the methodology used to extract valid data from THIN used for the analyses in **Chapter 5**. Methods are described for identifying the Acceptable Electronic BCSP (AEB) date for each GP practice receiving electronic notifications in THIN. Examples of the strategies developed to extract additional health data (AHD) variables are described with examples. Read code list development and definitions of variables used for the study in **Chapter 5** are presented along with the methods used to derive drug code lists. This chapter can be used in conjunction with **Chapter 5** to supplement further detail behind extracting a valid dataset for analysis and to ensure results are generalizable to the English BCSP population.

1.1 Structure of the THIN Database and coding information

Along with electronic GP record databases such as the Clinical Practice Research Datalink (CRPD) (formerly General Practice Research Database - GPRD) and QRESEARCH, THIN is a database of anonymised GP records used for health and medical research. This database is derived from GP systems which use the Vision operating system and is provided for research by IMS Health (now known as IQVIA). THIN provides data for over 587 practices (>5% coverage of the UK) covering more than 12 million patients. The data are made available for research by GPs entering notes onto the computer system, this is then anonymised and sent for collation in a database. Researchers then submit a protocol to the Scientific Review Committee (SRC) administered by IMS Health and if ethical approval is granted, the data is made available for research.

THIN provides information on diagnoses, symptoms, prescriptions, laboratory tests and lifestyle factors across four standardised data files linked by patient ID.[1] The patient file includes information such as registration to the practice, age and sex. The medical file contains information on diagnoses and symptoms, and the therapy file covers drug prescriptions. The Additional Health Data (AHD) file contains a range of information including laboratory test results, anthropometric measurements, smoking status and vaccinations. Additional Information Services (AIS) can also be used to obtain information collected from GP or patient questionnaires and data from other sources.

The raw data behind electronic health records such as THIN is not in a form which is ready for research. The data required for analysis needs to be transformed and extracted using various coding systems.[2] These coding systems can be exploited for data analysis by producing code lists for outcomes and covariates to extract from the database as well as identifying and researching a population with a particular disease condition or exposure. There are three main types of lookup which can be produced for THIN; Read code lists for use in the medical file (and also used in the AHD file), Drug code lists for use in the therapy file, and AHD codes for use in the Additional Health Data file.

Read codes named after Dr James Read are used to record clinical information such as symptoms and diagnoses in the UK and are used in the THIN database. Read codes use a hierarchical system of recording which is not based on ICD (International Classification of Diseases) which is the case in much of Europe. There are around 100,000 alphanumeric Read codes used to record the clinical information, the main stems are provided in **Table 1** below.[3]

Stem	Description
0	Occupations
1	History & symptoms
2	Examination and signs
3	Diagnostic procedures
4	Laboratory procedures
5	Radiology & physics in medicine
6	Preventative procedures
7	Operations, procedures & sites
8	Other therapeutic procedures
9	Administration
A	Infectious and parasitic diseases
B	Neoplasms
C	Endocrine, nutrition, metabolic and immunity disorders
D	Diseases of blood and blood forming organs
E	Mental disorders
F	Nervous system and sense organ diseases
G	Circulatory system diseases
H	Respiratory system diseases
J	Digestive system diseases
K	Genitourinary system diseases
L	Complications of pregnancy, childbirth and the puerperium
M	Skin & subcutaneous tissue diseases
N	Musculoskeletal and connective tissue diseases
P	Congenital anomalies
Q	Perinatal conditions
R	Symptoms, signs and ill-defined conditions
S	Injury & poisoning
T	Causes of injury and poisoning
U	External causes of morbidity and mortality
Z	Unspecified conditions

Table 1: The main stems of Read code classification. Taken from Davé and Petersen[3]

Prescriptions are linked to the British National Formulary (BNF) Chapter codes and can also be identified using their generic drug names. They are recorded in the THIN database as encrypted Multilex codes.[3] Health systems in the UK are also transitioning over to SNOMED CT which is an international coding system allowing greater research opportunities and analysis of health data.[4]

The AHD file contains information on lab test results, lifestyle data, immunisations and death data.[1] Each record has an AHD code which defines the area of interest (for example, smoking, alcohol, lab test results) and then further information is obtained from other data fields including a 'medcode' field (which contains Read codes) adding an additional layer of

information. Lab test results for example have the numerical value, units and reference values used by the lab. Pathology labs in hospitals send electronic results through to GP practices using the NHS Spine and Pathology Messaging Implementation Project Messaging (PMIP) to automatically populate patient records.[5]

GP practices have electronic links with the bowel cancer screening system using the same system that pathology labs use to send electronic results. Therefore, FOBT results and whether someone has participated in the programme are also available in the database. The AHD file stores the electronic BCSP notifications in THIN.

1.2 Bowel Cancer Screening Programme Notifications

The Bowel Cancer Screening Programme in England was rolled out in July 2006 and had national coverage by January 2010. Screening was offered to 60 to 69 year olds initially and extended to age 74 in 2010. The different hubs became involved in the screening programmes at different time points; Midlands and North West Hub started in July 2006, the Southern Hub September 2006, the London Hub October 2006, the North East Hub February 2007 and the Eastern Hub March 2007 (Steve Smith, personal communication). The Norwich screening centre started in July 2006. Although they are officially part of the Eastern Hub they started off independently for the first year and were subsequently transferred to the Eastern hub in July 2007 (Steve Smith, personal communication).

The different screening programmes offered in the UK use different age ranges, code lists as well as screening systems (Scotland, Ireland, England, Wales, and Isle of Man). For example, Scotland's IT system for colorectal cancer screening is called BoSS (Bowel Screening System).

The Read codes currently used by the English BCSP for electronic notifications to primary care are listed in **Table 2** and those by the Scottish Bowel Cancer Screening Programme in **Table 3**. In October 2013, Read Codes for Bowel Scope Screening were introduced (**Table 4**).

Read Code	Description
6866	Bowel cancer screening programme: faecal occult blood result
6867	Bowel cancer screening programme faecal occult blood testing kit spoilt
686A	Bowel cancer screening programme faecal occult blood test normal
686B	Bowel cancer screening programme faecal occult blood test abnormal
686C	Bowel cancer screening programme faecal occult blood testing incomplete participation
9Ow2	No response to bowel cancer screening programme invitation

Table 2: Set of Read codes used by the NHS BCSP in England to record colorectal cancer screening activity

Read Code	Description
686A	Bowel cancer screening programme faecal occult blood test normal
686B	Bowel cancer screening programme faecal occult blood test abnormal
6867	Bowel cancer screening programme faecal occult blood testing kit spoilt
68W2	Bowel Cancer Screening Programme – include reason of 'Ineligible'
9Ow3	Bowel cancer screening programme faecal occult blood testing incomplete participation
8IA3	Bowel Cancer Screening Declined
9Ow2	No response to bowel cancer screening programme invitation
66W2	Bowel Cancer Screening Programme - include reason of 'Non-Responder'

Table 3: Set of Read codes used by the Scottish Bowel Cancer Screening Programme to record colorectal cancer screening activity.

Read Code	Description
68W20	Bowel Cancer Screening Programme bowel scope screening test (Term 30 description is BCSP bowel scope screen test)
68W21	Bowel scope (flexible-sigmoidoscopy) screen: normal - no further action
68W2C	Bowel scope (flexible-sigmoidoscopy) screen: incidental findings
68W23	Bowel scope (flexible-sigmoidoscopy) screen: referred for colonoscopy
68W24	Bowel scope (flexible-sigmoidoscopy) screen: cancer detected
68W27	Bowel scope (flexible-sigmoidoscopy) screening invitation declined
68W28	Bowel scope (flexible-sigmoidoscopy) screening invitation: did not respond
68W29	Bowel scope (flexible-sigmoidoscopy) appointment: did not attend
68W2A	Bowel scope (flexible-sigmoidoscopy) screening: attended but not screened
68W2B	Bowel scope (flexible-sigmoidoscopy) screening invitation: unsuitable at this time

Table 4: Read code list for Bowel Scope Screening in the NHS BCSP in England

Since 2010, the Bowel Cancer Screening System (BCSS) has sent results of the FOBT electronically to GP practices who have opted into this service using the same system as the Pathology Messaging Implementation Programme (PMIP).[6] This service is also used by pathology labs to send laboratory results to primary care. The gFOBT results from screening are sent once daily overnight to the GP practices in batches.

1.3 Reproducible Research using THIN and other electronic GP databases

Due to the level of detail which can be provided by THIN, each study must define the variables required in terms of Read Codes used for symptoms and diagnoses. For instance, the Read codes used for the diagnosis of colorectal cancer may only include certain types based on study objectives. The Reporting of studies Conducted using Observational Routinely-collected health data (RECORD) statement recommends that code lists are provided with published studies either within the journal or linked to a data repository.[7] This statement consists of 13 items specific for observational studies which use routinely collected health data. More recently, it has been suggested that the actual code set engineering methods should also be provided for greater transparency to other researchers.[2] The inclusion of inappropriate codes and exclusion of appropriate ones is a recognised source of bias in studies of EHRs.[8]

When using previously developed Read code lists from other studies (e.g. Clinical Codes Data Repository: https://clinicalcodes.rss.mhs.man.ac.uk/), care must be taken that the list applies to the study objectives. In addition, different practices use different Read codes and new Read codes are introduced over time or drop out of use so it is important to update lists appropriately. In addition, the Quality and Outcomes Framework (QOF) introduced in 2004 has provided financial incentives for GPs to record certain clinical information.[9] For example, there have been QOF incentive schemes for health indicators such as smoking,[10] diabetes,[11] and severe mental illness[12] which have affected the pattern and quality of reporting over time. Other temporal changes which affect recording patterns are the introduction and changes to NICE guidelines.

The number of publications using EHR data in epidemiological and medical research has increased from around 80 in 2005 to more than 450 in 2015/2016.[8] This trend is set to continue with the move to electronic data and computerisation within the NHS. For example, a change in health policy has required that all healthcare trusts will only use electronic referrals from GPs by October 2018.[13] The methods for extracting data and using it for analysis therefore needs to be rigorous and reproducible. In addition, new methods are required to assimilate all this information and use it for health sciences research.

Furthermore, data assurance in terms of which dates to include in analyses based on different events in time, such as the introduction of Vision software/switch overs must be investigated for data quality assurance. The Acceptable Mortality Reporting (AMR) date

and Acceptable Computer Usage (ACU) date have been derived for this purpose. The THIN database also includes several quality indicators such as patient flags (Patflags) in the Patient file.

1.4 Quality Assurance Indicators Derived for THIN Studies

Previous data quality filters have been developed for use in THIN.[14] [15] The AMR date was developed to define the periods of acceptable mortality reporting for each practice in THIN.[14] Before this date, practices may not have routinely recorded deaths or de-registrations. Applying this external standard to the data provides validity to the data extracted and removes under-reporting of death. The AMR date is supplied by THIN and is the date at which mortality reporting matches that of the general UK population. Studies using THIN apply the AMR date to each practice to define the start of patient follow up. The study reported in **Chapter 5** to define the time period eligible for each practice used the latest of several dates including the AMR date to ensure validity of extracted data.

Another data quality filter which has been developed for use in THIN is the ACU date.[15] This date was developed to define periods of acceptable computer usage for each practice in THIN by using the criteria of one medical, one AHD and two therapy records per patient per year on average. When new software and computer systems were implemented in practices, there was a period of adaptation to the new systems by health care practitioners. As a result, diagnoses or other recordings could be affected following the adoption of this new technology. The authors suggest using the latest of the AMR and ACU date to define the start of patient follow up and improve data quality.[15]

Studies using recording/notification systems which have been implemented over time will therefore need additional quality filters to ensure recordings are valid and representative of the UK population. The Bowel Cancer Screening System (BCSS) was developed alongside the BCSP to record screening activity and results for patients eligible for screening. As described above, the screening programme was sequentially rolled out in July 2006 and practices were eligible to receive electronic notifications by opting into this service from the 1st April 2009. Before the adoption of this system, recording of BCSP FOBT results was input by practices on receipt of letters sent by the BCSP. Additional data on screening activity is also sent such as non-response, FOBT kit spoilt and incomplete participation. Correspondence with NHS Digital suggests that 88% of all GP practices now receive electronic GP communications (Stephen Halloran, personal communication). The remaining

practices may not serve the population of interest (University practices for example) or opted for paper reports which may give additional information (Stephen Halloran, personal communication). There are discussions of improving this interconnectivity between GP records and screening records. The electronic notifications are sent using the same system as the PMIP. There is therefore also scope for future research to define acceptable recording dates for lab test results received from pathology for the same reasons described above for the FOBT.

1.5 Rationale

In order to derive valid data for the study reported in **Chapter 5**; Read code and drug code lists along with AHD variables needed to be derived and tailored specifically to the study objectives to ensure that results are generalizable to the English BCSP population. Further to this, a date of acceptable electronic BCSP notifications was required to define practice inclusion and the start of patient follow up.

The objective of this chapter was to describe the methodology used to derive;

i) AEB date for each practice
ii) AHD variable extraction with examples of FOBT screening outcomes and haemoglobin concentration
iii) Read code lists with an example of bowel cancer diagnosis
iv) Drug code lists with an example of laxatives

This chapter can be used in conjunction with **Chapter 5** to supplement further detail behind extracting the dataset for analysis.

2.0 METHODS & RESULTS

Stata 14.0 was used to derive an AEB date, for AHD variable extraction methods and to compile Read code and Drug code lists for data extraction.

3.0 Developing Methodology to define an AEB (Acceptable electronic BCSP) date

3.1 Methods: Defining an AEB Date

To derive an AEB date for each practice opted into receiving electronic BCSP notifications, the incidence of BCSP FOBT results for people aged 60-74 were examined. This revealed a time-point at which the practice started to receive the electronic notifications from the bowel cancer screening system. The uptake rate for bowel cancer screening is around 50%[16] and people are invited biennially. As a result, it is expected that approximately 25% of those aged 60-74 will have a FOBT result each year. Consequently, the time-point at which the practice starts to receive FOBT results can be identified by examining the date on which the frequency of Bowel Cancer Screening results (using the nationally agreed Read codes for reporting screening outcomes) in people aged over 60 rises to an expected monthly rate. The FOBT results from this date onwards will predominately be the electronically received notifications from the BCSP.

3.1.1 Setting up a numerator/denominator for the AEB date

THIN (Version: May 2016) was used to derive the AEB date. Only practices in England were considered for inclusion due to the differing Read codes used for recording results and different bowel cancer screening systems across the regions. Some preliminary analyses revealed that regions differ in the frequency and use of codes. Here the focus is on codes which are used by the English Bowel Cancer Screening Programme (BCSP) and bowel cancer screening system (BCSS).

The frequency of BCSP notifications by number of patients registered in a practice age 60-74 were investigated. The frequency of patients registered in a practice aged 60-74 for each month (and for each primary care practice in THIN) were obtained for use as a denominator. The numerator consisted of the frequency of electronic BCSP notifications

received per month. This rate was multiplied by 1000 before plotting on a line graph for each practice in England.

An expected rate for each practice was also generated. It is expected that around 25% of those aged 60-74 will have a FOBT result each year (based on 50% uptake and biennially), this was divided by the number of patients registered a month and multiplied by 1000 for comparison and for overlaying on the line graph (monthly expected rate).

The definitions of variables used to derive the AEB date are shown in **Table 5**.

3.1.2 Denominator

The denominator consisted of patients registered in that practice at the beginning of each month aged 60-74 (operationalised as 59-75 to ensure the recorded birthdays on THIN cover the population of interest).

An eligibility start date variable was generated which was the maximum of the registration date, date aged 59, the Vision date or the 1st April 2009 (which was when practices had the ability to opt into electronic screening notifications). An eligibility end date variable was also generated which consisted of the minimum of the transfer date, date aged 75 and the collection date. Any records where the eligibility end date was less than or equal to the eligibility start date were removed from the analysis.

The command 'stset' in Stata was then used to allow splitting of the data with the eligibility end date as the exit variable, by each patient ID and eligibility start date as the entry variable. This data was then split by month to determine the number of notifications/people registered a month. The patient-level data was then collapsed to provide a count of registered patients by practice and month for the denominator.

3.1.3 Numerator

Both the AHD file and Medical file were used to derive the Bowel Cancer Screening codes of interest – the method to extract BCSP records is detailed in **Section 4**. The AHD file contains the electronic BCSP notifications and previously recorded paper based results. The medcode file could have also been used to record paper copies of results during a consultation and was also used to derive BCSP codes. This ensured all BCSP records were extracted. Any duplicate records were dropped from analysis to ensure that the numerator was not inflated. The data were made so that it was one line per patient by dropping any duplicates in terms of patient ID, Read code, event date and practice ID. To ensure that the event date is within the period of interest, any dates before 1st April 2009 (which is the date FOBT codes were released) were excluded.

Generated Variable	Description
Numerator	The frequency of electronic BCSP notifications received per month in each practice Using both AHD file and Medcode file (removing duplicate records to ensure one row per person event) Restricting to those who were aged 60-74 for the screening event (electronic notification) Within the period of interest (i.e. after 01 April 2009) Practices in England
Eligibility start (to create denominator)	The maximum of the registration date, date aged 59, the vision date or 1st April 2009 (which was when the first practices had the ability to opt into electronic screening notifications)
Eligibility end (to create denominator)	The minimum of the transfer date, date aged 75 and the collection date
Denominator	Patients registered in that practice at the beginning of that month aged 60-74 Practices in England
Actual monthly rate by practice (based on number of monthly BCSP notifications over patients registered per month)	actual_rate = (numerator/patients_registered)*1000
Expected monthly rate by practice (based on patients registered per month)	expected_monthly_rate = (((patients_registered*0.25)/12)/patients_registered)*1000 Average 20.83

Table 5: Variables created to set up a numerator/denominator for the AEB date

3.1.4 Initial sort of practices for inclusion/exclusion – visual review

Line graphs with the expected monthly rate of electronic notifications and the actual monthly rate by practice were produced along with a locally weighted scatterplot smoothing (LOWESS) line and visually examined. The LOWESS was plotted to assist with identifying the point at which practices receive electronic notifications as sometimes the recordings appeared intermittent.

Inclusion and exclusion criteria for GP practices and their BCSP electronic notification patterns were derived and agreed upon by the reviewers (**Table 6**). An initial sort was carried out by two reviewers (JC, TM) to decide which practices to include for date assignment or exclude based on the exclusion criteria below. A percentage agreement was calculated. A third reviewer (RR) resolved any disagreements so a consensus was reached.

Exclusion criteria	Inclusion criteria
1. Too short a period to be useful or to judge stability (exclude)	1. There is a clear sharp increase in the rate of electronic notifications received (include)
2. If the rate is too low (exclude)	2. A screening start date can be derived from the peaks (include)
3. If the rate is 0 (exclude)	
4. Intermittent small peaks, no distinct rise (exclude)	

Table 6: Inclusion and exclusion criteria for GP practices and their BCSP electronic notification patterns.

3.1.5 Assigning the AEB – visual review

The month and year were recorded for each practice by one reviewer (JC) and a consensus meeting was carried out with a second reviewer (TM) to discuss interpretation of the start date (the change point) based on the rules in **Table 7** below. A third reviewer (RR) in the consensus meeting was consulted if there was disagreement between the two reviewers (JC, TM).

Rules for visual interpretation of the start date:
1. Assess the first peak and take the start date as one month after the first peak to ensure consistency. 2. If there are two or more distinct peaks, go by the second or last peak to ensure consistent results. 3. For multiple peaks, if there is a period of several months at a low/zero rate then assess the next significant increase in notifications.

Table 7: Rules for visual interpretation of the AEB start date for each included practice.

3.2 Results: Defining an AEB Date

The THIN database version used for this analysis was May 2016 where data were recorded for 455 practices. The initial sort of these practices by the first reviewer (JC) gave 353 practices for inclusion and 102 for exclusion. The second reviewer gave 363 for inclusion and 102 for exclusion giving 97.8% (445/455) agreement. The consensus meeting with the 3rd reviewer gave the final 363 for inclusion and 92 practices for exclusion. Eighty per cent of practices were therefore investigated for an AEB date. Some practices may have been excluded because they had too short a duration before they stop contributing to THIN. In addition, some practices were excluded when it was not clear where the electronic notifications commenced (no distinct rise).

3.2.1 Included Practices

When deriving the AEB date for the included practices, the majority had a straightforward visual increase in the monthly rate of BCSP notifications over patients registered in the practice (263/363 practices had this electronic notification pattern). This pattern is shown in **Figure 1** where the blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time.

The different patterns in electronic notifications seen by the practices can be influenced by practice behaviour, patient behaviour, the sending of lab results and the change/update in IT Systems such as Vision software. Patient behaviour in completing the test at home can cause delay (for example over Christmas and other holidays). Practices may have received results by letter and then changed to electronic notifications or updated IT systems at different time points leading to delays in the receipt of results.

Figure 2 shows an example where there is no distinct increase in rate but the AEB date can still be derived. A big spike could mean there was a backlog of inputting/receiving results. **Figure 3** shows an example where there is an initial peak followed by a few dips. **Figure 4** shows a clear increase but then the practice stops contributing to THIN. **Figure 5** shows an intermittent peaking pattern with a small peak followed by a larger one; this scenario could be two practices merging together for example. **Figure 6** shows many dips but also a

general upwards trend; the dips could be the sending of results in batches to practices by the lab.

Other patterns showed peaks gradually decreasing over time which could reflect the movement of patients. Practices with prolonged gaps were still considered for inclusion since this could reflect a breakdown in computer software. All types of practices were considered so bias was not introduced. The derived AEB date for the included examples are shown in **Table 8.**

Practice ID (pseudonymised)	AEB Date
0399	01/09/2012
0084	01/07/2009
0194	01/12/2011
0165	01/10/2013
0388	01/11/2010
0451	01/04/2012

Table 8: Examples of included practices and corresponding AEB Date.

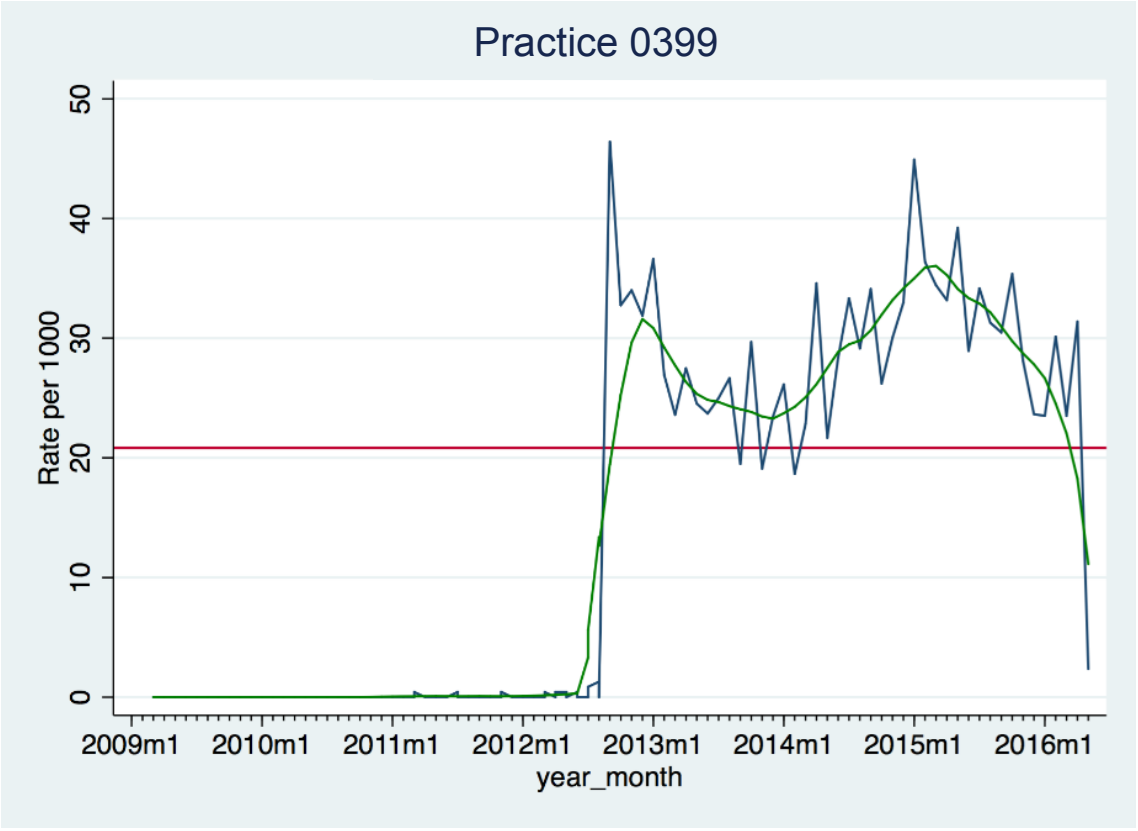


Figure 1: Example of a practice with a clear visual increase in the start of receiving electronic BCSP notifications. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st September 2012.

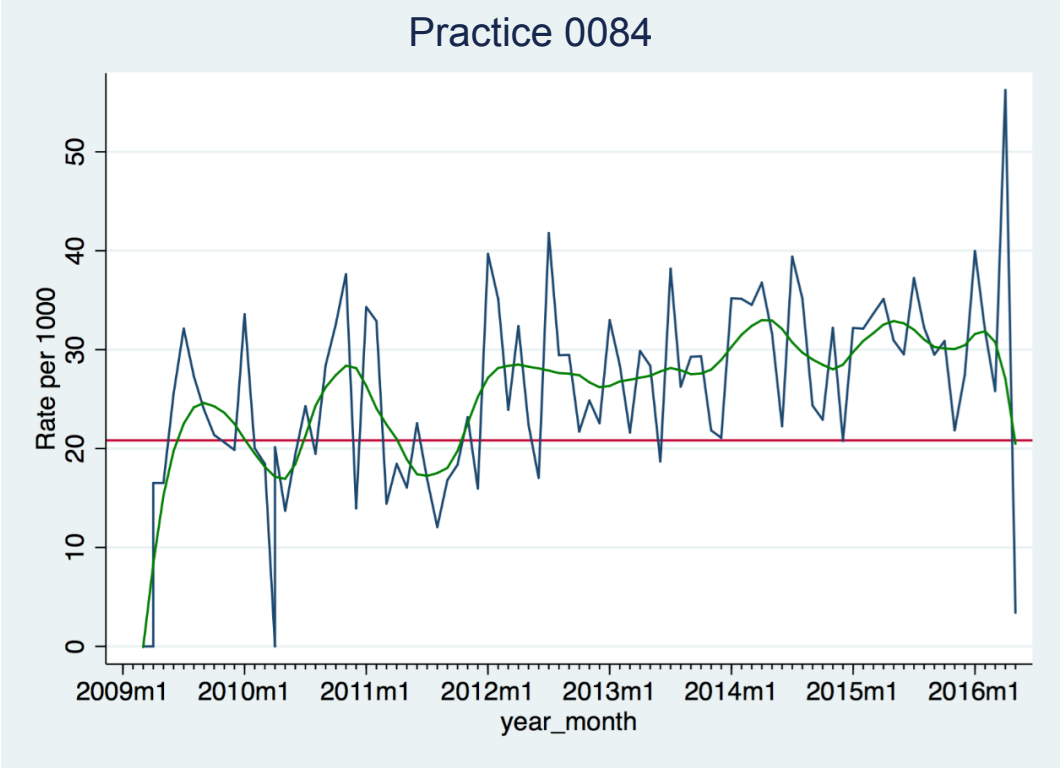


Figure 2: Example of a practice where there is no distinct increase in rate but the AEB date can still be derived. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st July 2009.

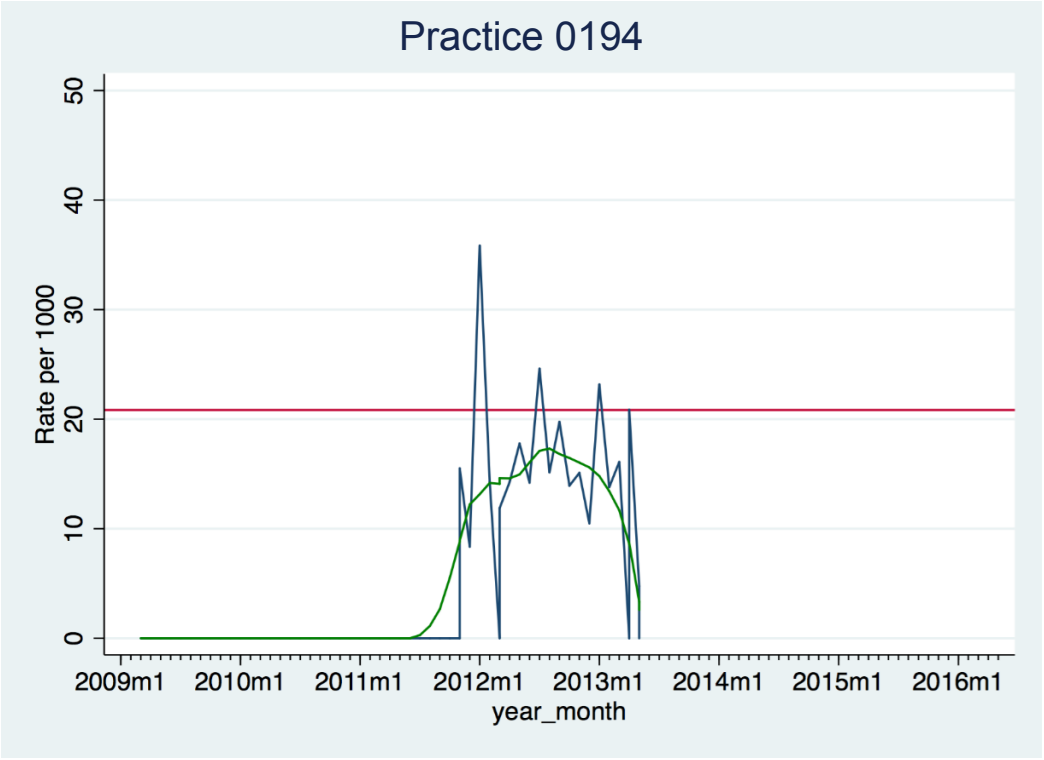


Figure 3: Example of a practice where an initial peak followed by a few dips. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st December 2011.

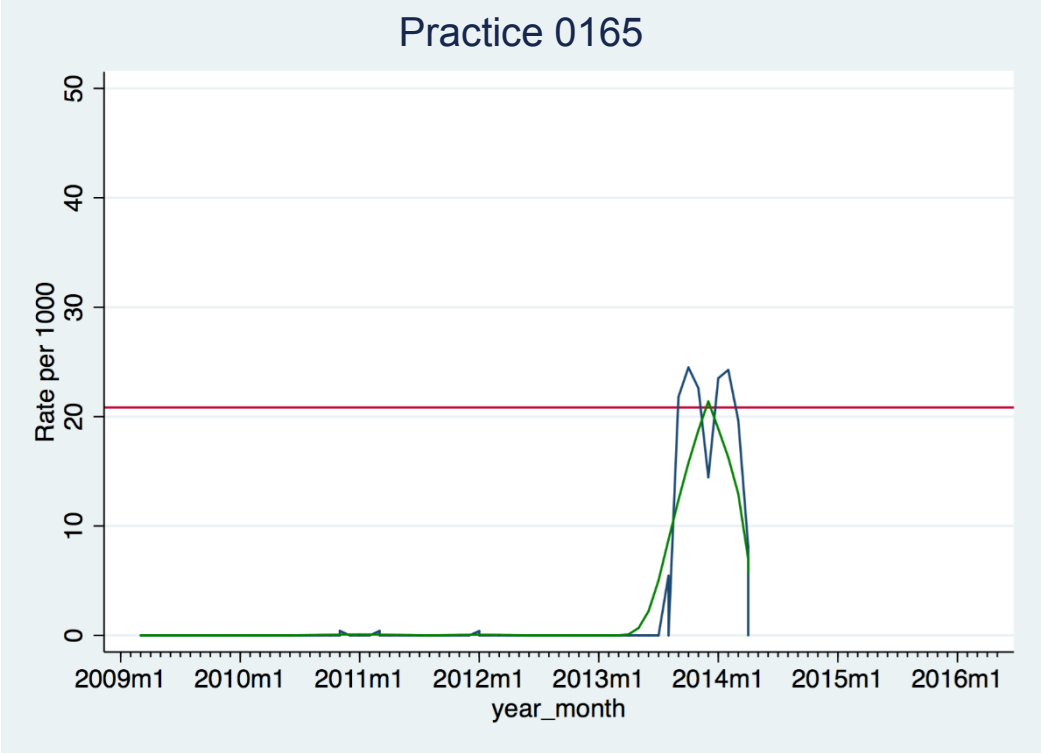


Figure 4: Example of a practice where a visual peak is observed but then the practice stops contributing to THIN or receiving FOBT results. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st October 2013.

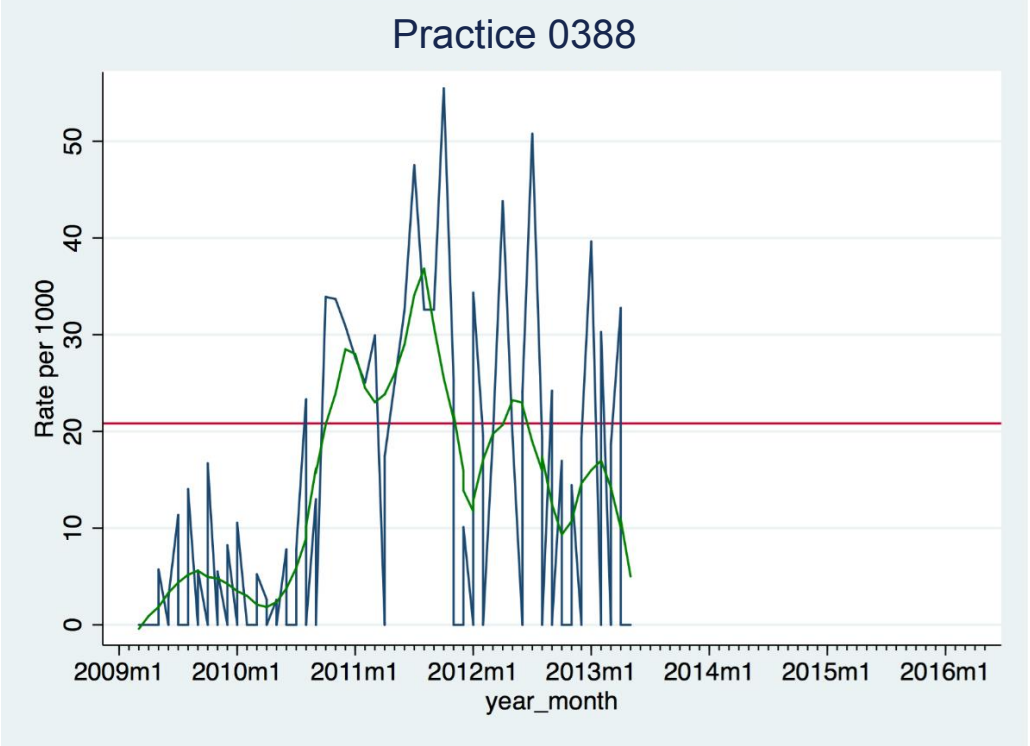


Figure 5: Example of a practice where there is an intermittent peaking pattern with a small peak followed by a larger one. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st November 2010.

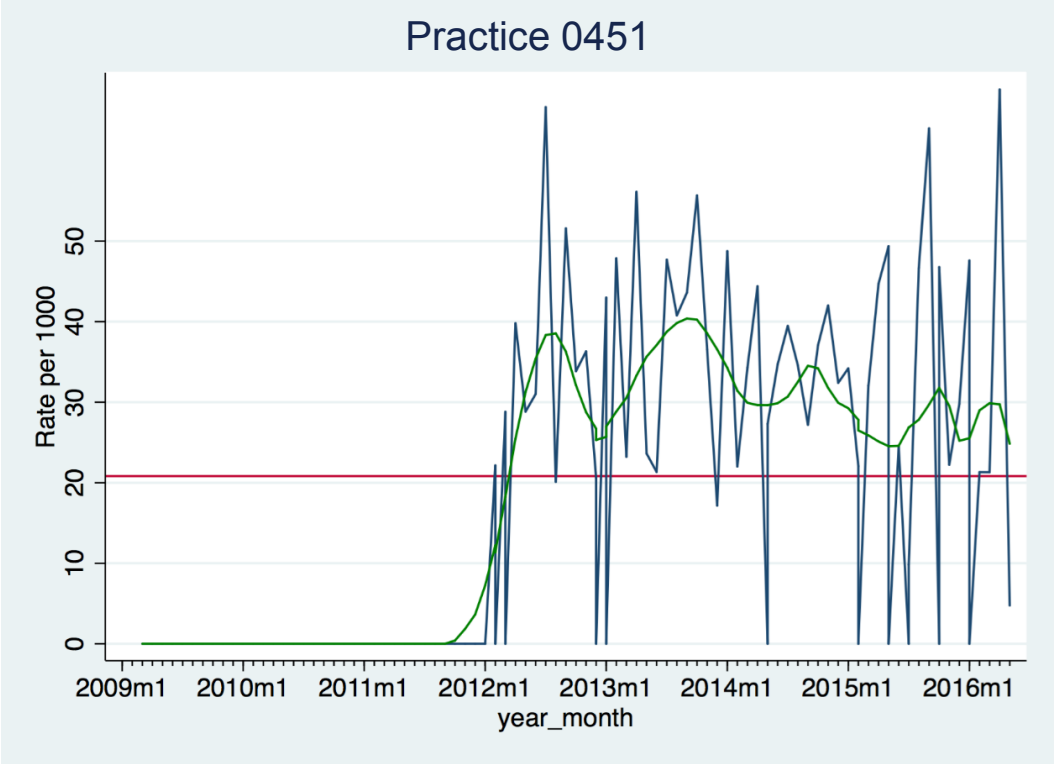


Figure 6: Example of a practice which shows many dips but there is also a general upwards trend. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time. The AEB date for this practice was assessed as the 1st April 2012.

3.2.2 Excluded Practices

Examples of excluded practices are included below. **Figure 7** shows a practice with very small peaks which suggests electronic notifications may not be being received. Alternatively, the practice could be very small or serve very few numbers in the age range of interest. **Figure 8** presents a practice where there is a nil rate and therefore appears there is no recording of electronic notification activity (practice may not serve the age range of interest). Finally, **Figure 9** shows a practice where there is an extremely low rate of notifications received intermittently. This could be due to paper copies of results perhaps being received by a smaller practice.

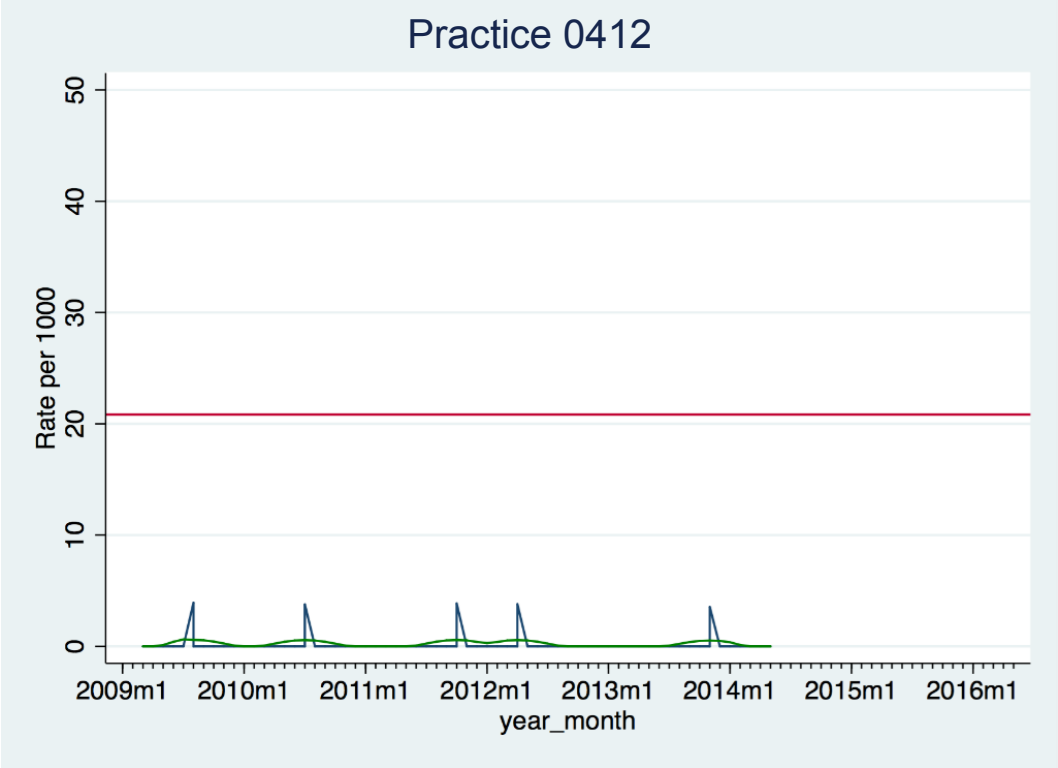


Figure 7: Example of a practice which is excluded. There are small peaks in notifications. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time.

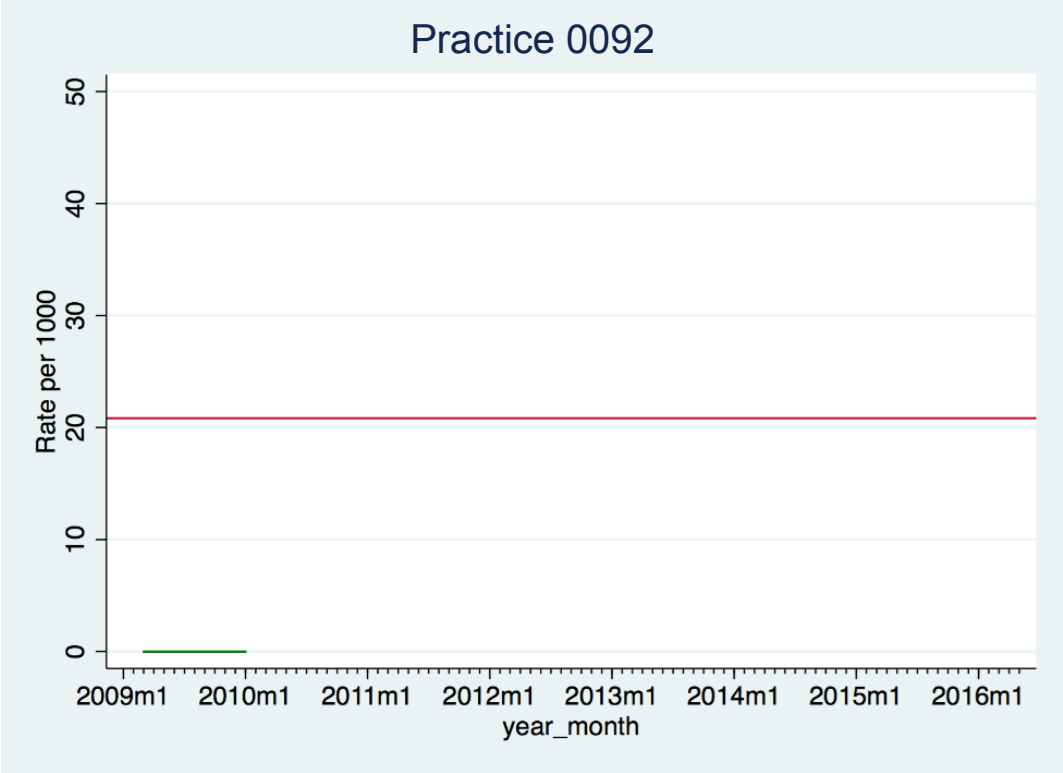


Figure 8: Example of a practice which is excluded. There is a nil rate and appears to be no recording of screening notification activity. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time.

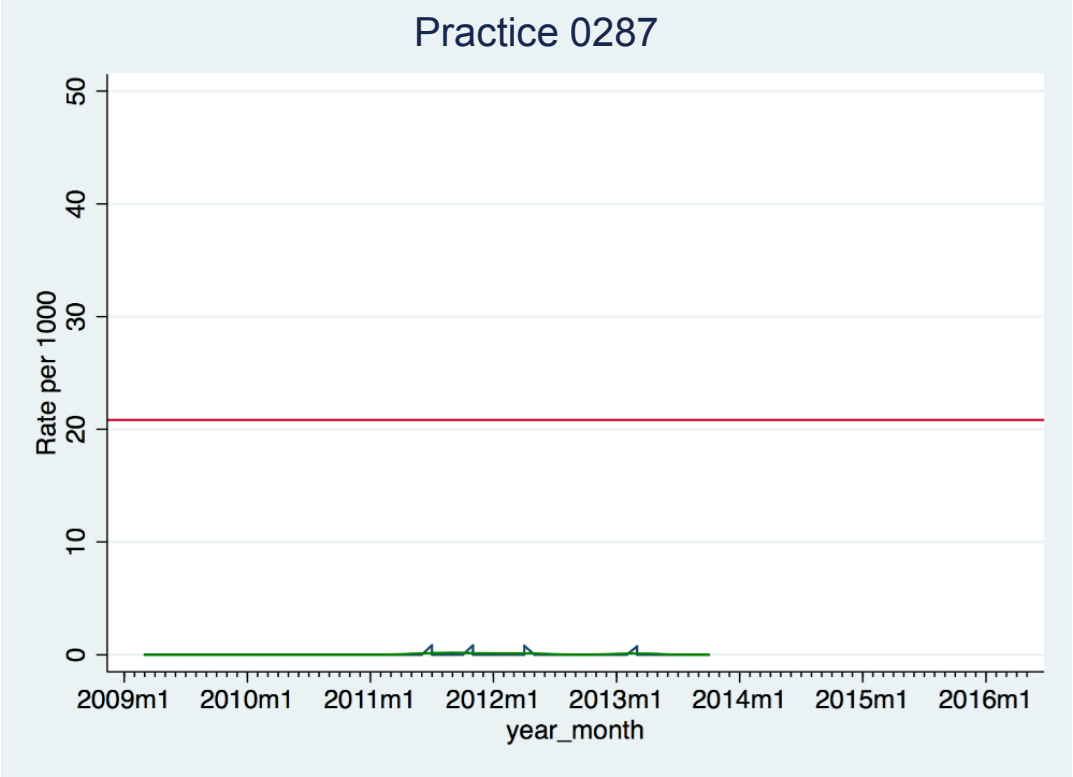


Figure 9: Example of a practice which is excluded. There are irregular very small peaks in notifications and then the practice stops contributing to THIN. The blue line shows the actual rate of BCSP notifications, the red horizontal line displays the expected rate of notifications (20.83). The green line is the LOWESS line used to assess the overall trend in the change of electronic notifications over time.

The AEB date for each included practice is given in **Appendix 1**. These dates were used when deriving the data from THIN for the study reported in **Chapter 5** as a data quality filter and to define the period of interest. Any practices without an AEB date where excluded from the analysis, and when defining the practice start dates this was the latest of the AMR, AEB date and Vision date plus one year.

4.0 Devising a method to extract AHD File Variables

The Additional Health Data file contains information on lab test results, lifestyle data, immunisations and death data.[1] Each record has an AHD code which defines the area of interest (for example, smoking, alcohol, lab test results). Further information is obtained from data1 to data6 fields including a medcode field (which contains Read codes) adding an additional layer of information. For example, if the AHD code is for smoking, the other columns will give details such as whether someone is a current smoker, number of cigarettes a day, start date and stop date if applicable. Likewise with lab test results such as haemoglobin concentrations, the other fields will give additional information such as the numerical concentration, the units used to record this and the reference values used by labs which gives an indication of the validity of the result. AHD codes have a lookup in THIN to obtain the description.

Laboratory measurements (such as blood test results) need to consider the different units which are recorded/used by pathology labs in England, the reference distribution for haemoglobin concentration (to ensure the correct units are used) and plausible minimum and maximum values from an external data source to remove potential outliers.

For the study reported in **Chapter 5,** methods to extract variables from the AHD file were developed for the following: Blood group, family history of gastrointestinal cancer, FOBT primary care outcome, FOBT screening outcome, haemoglobin concentration, mean cell volume, ferritin concentration and platelet count.

The methods for both the FOBT screening outcome and haemoglobin concentration are given as examples below. Documents for the methods for the other variables listed above are available from the author.

4.1 Methods: Development of a method to identify FOBT screening outcomes from the AHD file in THIN

Since 2010, the Bowel Cancer Screening System (BCSS) has been able to send the results of the FOBT electronically to GP practices using the same system as the PMIP service.[17] The results are sent once daily overnight (every working day) to the GP practices in batches which are queued that day. Previously, since 2006 results were sent by hardcopy letter from the hub to the GP practices where the practice was responsible for recording the result in the patient records. The timing of hardcopy letters sent is determined by each

hub.[17] There are a set of Read codes which have been generated for the English BCSP which differentiate from the test results of FOBTs requested and performed through primary care.

This section describes the development of a method to extract FOBT screening outcomes from the AHD table in the THIN database. This enables the BCSP history to be extracted for an individual.

The method was an iterative process which systematically reviewed the different combinations of Read codes and value labels in the AHD data observed in a THIN 1% sample, initially restricting by the ahdcode of interest (FOBT). The AHD data was then searched for any additional BCSP codes not restricting by ahdcode.

Medcodes which gave a definitive outcome/result such as, 'BCSP FOB test normal' (686A.00) could be extracted without further information required from the data1-data6 columns (in this case data4 was the column populated). This was the case for 94.53% of the data under the selected ahdcode 'FOBT' (4645/4914). The generic BCSP code on the other hand ('Bowel cancer screening programme: faecal occult blood result' (6866.00)) needed to be combined with one of the following ahdcodes from data4 to obtain a definitive outcome; P/N001 (negative), P/N002 (positive), PTH005 (normal), PTH010 (abnormal). Using this method, 3.03% of the data could be utilized this way (149/4914). Generic codes without further information could be dropped (2.44% of the data 120/4914). A further 28 codes were identified when not restricting by ahdcode (all 9Ow2.00), this code did not need to be combined with further data as it was a definitive outcome.

The output is a series of rules which will identify BCSP FOBT screening outcomes for all patients in the THIN database. Further work may be required to exclude individual influential outlying results from specific analyses.

Data Source

This analysis used a 1% random sample of all patients from THIN1601 as its data source.

Analysis Tools

Stata version 14 was used to tabulate the frequency of Read code and value label combinations. The main command used to describe the data were 'tabulate'.

Procedure

1. Identification of ahdcode(s) likely to contain BCSP FOBT Screening Outcomes.
2. Restriction of medcodes (i.e. Read codes) to those associated with the BCSP.
3. Tabulation of Read code and value label combinations by frequency.
4. Search for BCSP FOBT Read codes which may have been recorded under another ahdcode.
5. Description of the final method: a set of rules to be applied to identify BCSP FOBT Screening Outcomes.
6. Example code for these rules in Stata

4.2 Results: Development of a method to identify FOBT screening outcome from the AHD file in THIN

4.2.1 Identification of ahdcode(s) likely to contain BCSP FOBT Screening Outcomes.

The system lookup ahdcodes.dta was used to identify codes of interest to search in the ahd.dta (AHD file). The following ahd code was identified:

Datafile	Ahdcode	Description
TEST	1001400080	Faecal Occult Blood

Table 9: Ahdcode identified from the system lookup table ahdcodes.dta.

AHD codes were searched for the following key terms using regular expressions and scanning the data manually for: faecal occult blood test, FIT, Faecal immunochemical test, BCSP and screening.

4.2.2 Restriction of medcodes to those associated with the BCSP

The medcodes associated with this ahdcode were summarised and any codes which were not associated with the BCSP were excluded from the analysis.

Medcode	Description	Frequency
6867.00	BCSP FOB testing kit spoilt	1
~~4793.00~~	~~Faecal occult blood: trace~~	1
~~8IA3.00~~	~~Bowel cancer screening declined~~	3
~~68W2.00~~	~~Bowel cancer screening programme~~	6
~~4795.00~~	~~Serial faecal occult blood normal~~	47
~~4794.00~~	~~Faecal occult blood: positive~~	47
~~479Z.00~~	~~Faecal occult blood NOS~~	71
~~4791.00~~	~~Faecal occult blood requested~~	79
686C.00	BCSP FOB tst incmplt particpt	87
686B.00	BCSP FOB test abnormal	92
~~479..11~~	~~Faeces occult blood test~~	110
6866.00	Bowel cancer screening programme: faecal occult blood result	269
~~4792.00~~	~~Faecal occult blood: negative~~	564
686A.00	BCSP FOB test normal	4465
~~479..00~~	~~Faecal occult blood test~~	5669

Table 10: Frequency of medcodes recorded under the Faecal Occult Blood 1001400080 AHD Code. The codes which have been scored out are those which were not related to the BCSP for FOBT results.

The medcodes which were not related to the BCSP could be excluded from the analysis.

These included: 4795.00, 4794.00, 479Z.00, 4791.00, 4793.00, 479..11, 4792.00, 479..00.

68W2.00 relates to Bowel Scope Screening and was removed along with 8IA3.00 which is not one of the electronically sent codes from the BCSP. The other codes relate to FOBTs requested through primary care.

6866.00 can be combined with other information to go within other categories (whether a positive or negative FOBT result) and is classed as a ‘generic code’. All other medcodes give a definitive outcome and can be used on their own without additional value labels (data1-data6).

Medcode	Description	Frequency
6867.00	BCSP FOB testing kit spoilt	1
686C.00	BCSP FOB tst incmplt participt	87
686B.00	BCSP FOB test abnormal	92
6866.00	Bowel cancer screening programme: faecal occult blood result	269
686A.00	BCSP FOB test normal	4465

Table 11: Medcodes of interest for FOBT Screening Outcomes after excluding those codes associated with primary care.

4.2.3 Tabulation of Read code and value label combinations by frequency

From scanning through the data1-data6 columns in the ahd.dta file, only data4 was populated with a value label.

The units in the data4 column were summarised **Table 12.**

Data4	Lookupdesc	Frequency
P/N002	Positive	17
PTH010	Abnormal	71
P/N001	Negative	564
		930
PTH005	Normal	3332

Table 12: Value label descriptions for data derived from the AHD file relating to the 'Faecal Occult Blood' ahd code.

Where there were no value labels this probably relates to the medcodes which do not need/have any further information. They give a definitive result on their own e.g. BCSP FOB test normal (686A.00).

The combinations of medcodes and value labels from data4 were summarised to determine how the data was generally being recorded (**Table 13**):

data4	medcode	Frequency	Description of codes
P/N002	686C.00	1	Positive/BCSP FOB tst incmplt particip
	6867.00	1	BCSP FOB testing kit spoilt
P/N002	6866.00	3	Positive/Bowel cancer screening programme: faecal occult
	686B.00	8	BCSP FOB test abnormal
P/N002	686B.00	13	Positive/BCSP FOB test abnormal
PTH010	686B.00	71	Abnormal/BCSP FOB test abnormal
	686C.00	86	BCSP FOB tst incmplt participt
	6866.00	120	Bowel cancer screening programme: faecal occult blood
P/N001	6866.00	146	Negative/Bowel cancer screening programme: faecal occult
P/N001	686A.00	418	Negative/BCSP FOB test normal
	686A.00	715	BCSP FOB test normal
PTH005	686A.00	3332	Normal/ BCSP FOB test normal
	Total	4914	

Table 13: Medcode and value label combination frequencies

As identified above, 6866.00 (generic FOBT code) can be combined with the value labels in data4 (P/N001 (negative), P/N002 (positive), PTH005 (normal), PTH010 (abnormal)) to give definitive outcomes. All other medcodes can be used on their own as they give a definitive outcome.

4.2.4 Search for BCSP FOBT Read codes which may have been recorded under another ahdcode.

In some instances, some of the screening codes (medcodes) may have been recorded under another ahdcode (other than the one relating to 'Faecal Occult Blood'). The medcodes related to BCSP FOBT Screening Outcomes (identified from Read code lookups) were therefore searched in the whole 1% ahd file (i.e. not restricted by ahdcode).

The following Read codes were searched for in the ahd data:

6866. - Bowel cancer screening programme: faecal occult blood result
6867. - Bowel cancer screening programme faecal occult blood testing kit spoilt
686A. - Bowel cancer screening programme faecal occult blood test normal
686B. - Bowel cancer screening programme faecal occult blood test abnormal
686C. - Bowel cancer screening programme faecal occult blood testing incomplete participation
9Ow2. - No response to bowel cancer screening programme invitation

By tabulating the ahdcode, two other ahdcodes were identified (1001400153 and 1001400329) which are used to record BCSP outcomes. These were not frequently used:

ahdcode	Freq	Percent	Cumulative %
1001400080	4,914	99.43	99.43
1001400153	8	0.16	99.60
1001400329	20	0.40	100.00
Total	4,942	100.00	

Table 14: AHD codes from all the Read codes related to BCSP FOBT Screening Outcome.

Tabulating the medcodes, identified that there was an additional medcode not covered by the ahdcode for 'Faecal Occult Blood'.

Medcode	Freq.	Percent	Cumulative %
6866.00	269	5.43	5.43
6867.00	1	0.02	5.45
686A.00	4,465	90.18	95.64
686B.00	92	1.86	97.50
686C.00	87	1.76	99.25
9Ow2.00	28	0.57	100.00
Total	4,942	100.00	

Table 15: All medcodes from the ahd data file not restricted by ahd code.

The additional 28 codes from the other ahdcodes are related to the 90w2.00 Read code (No response to bowel cancer screening programme invitation). A rule can be included to search the whole of the ahd file for this Read code or from the additional ahdcodes (1001400153, 1001400329). It also has a 'definitive outcome' so does not need to be combined with further data/information from data4.

The uptake of the gFOBT is between 50-60% and therefore a similar proportion of 'no response to bowel cancer screening programme invitation' codes are expected as the proportion of FOBT result codes derived from THIN. However, this is not the case from the data derived. A significantly smaller number of 'no response codes' are seen in this dataset which suggests that this code is not being used for this purpose. This could be because it is mainly the test result (positive/negative) that the GP has responsibility to record.

4.2.5 Description of the final method: a set of rules to be applied to identify BCSP FOBT Screening Outcomes.

Based on all of the above investigations the following rules can be used to extract data for patients with FOBT Screening Outcomes.

1. Select all records with ahdcode 1001400080 (Faecal Occult Blood)

2. Keep any records when the medcode is equal to the following:

"6866\.00|6867\.00|686A\.00|686B\.00|686C\.00|9Ow2\.00"

3. Definitive BCSP code can be a medcode on its own.

686A.00 – BCSP FOB test normal

686B.00 – BCSP FOB test abnormal

686C.00 – BCSP FOB tst incomplt participt

6867.00 – BCSP FOB testing kit spoilt

4. Generic BCSP code (6866.00) needs to be combined with a data4 ahdcode for a definitive outcome (P/N001 (negative), P/N002 (positive), PTH005 (normal), PTH010 (abnormal)).

6866.00 – Bowel cancer screening programme: faecal occult blood result

5. Recode the description of the above 'generic code plus data4 ahdcodes' to fall into one of the defined categories (BCSP FOB test normal/BCSP FOB test abnormal). See **Table 16**.

Category	medcode	Generic Combinations also classed under this category
BCSP FOB test normal	686A.00 Under AHD code 1001400080 (Faecal Occult Blood)	6866.00 – Bowel cancer screening programme: faecal occult blood result With ahdcode of: P/N001 Negative PTH005 Normal
BCSP FOB test abnormal	686B.00 Under AHD code 1001400080 (Faecal Occult Blood)	6866.00 – Bowel cancer screening programme: faecal occult blood result With ahdcode of: P/N002 Positive PTH010 Abnormal

Table 16: Summary of combinations to include to identify BCSP FOBT Screening Outcomes.

6. Drop records with a Generic BCSP code but no further codes/information recorded under data4 (6866.00 on its own).

7. Include 9Ow2.00 under any ahdcode - No response to bowel cancer screening programme invitation (this does not need to be combined with anything else as it gives a 'definitive' outcome).

8. Drop if the event date is missing for any of these outcomes as the FOBT outcome date is required for the index date and to build a picture of screening history/analysis.

These rules are summarised in **Table 17**.

Summary:

Category	medcode	Generic Combinations also classed under this category
BCSP FOB test normal	686A.00 Under AHD code 1001400080 (Faecal Occult Blood)	6866.00 – Bowel cancer screening programme: faecal occult blood result With ahdcode of: P/N001 Negative PTH005 Normal
BCSP FOB test abnormal	686B.00 Under AHD code 1001400080 (Faecal Occult Blood)	6866.00 – Bowel cancer screening programme: faecal occult blood result With ahdcode of: P/N002 Positive PTH010 Abnormal
BCSP FOB tst incomplt participt	686C.00 Under AHD code 1001400080 (Faecal Occult Blood)	-
Bowel cancer screening declined	8IA3.00 Under AHD code 1001400080 (Faecal Occult Blood)	-
BCSP FOB testing kit spoilt	6867.00 Under AHD code 1001400080 (Faecal Occult Blood)	-
No response to bowel cancer screening programme invitation	9Ow2.00 Under any ahdcode	-

Table 17: Summary of the combinations of medcodes and data4 codes used to extract BCSP outcomes.

4.2.6 Example code for these rules

The example code for use in Stata using the rules above is provided in **Appendix 4**.

4.3 Methods: Development of a method to identify haemoglobin concentration values in THIN

Haemoglobin concentration (Hb) is a continuous value which has been most commonly recorded by pathology services as grams of haemoglobin per litre of blood (g/L) or grams of haemoglobin per decilitre of blood (g/dL). The Pathology Harmony Initiative[18] notified all UK laboratories to standardise the units for full blood counts including Haemoglobin level and MCHC to g/L in April 2012.[18] For example, 12 g/dL should now be reported as 120 g/L. UK laboratories then made this switch over the next year and ideally by 31st March 2013. This section describes the development of a method to extract valid Hb values from the AHD table in the THIN database.

The method was an iterative process which systematically reviewed the numeric distribution of Hb values for combinations of Read codes and value labels observed in a THIN 1% sample. These were compared with a reference distribution for Hb to assess if they matched that distribution, required conversion before use, or were unlikely to contain Hb values and therefore excluded. Plausible Hb minimum and maximum values from an external source were then applied to the data as a final step in the method.

Approximately 36.56% of Hb values associated with the final ahdcode of interest were Read coded or labelled differently than the reference distribution. This higher than expected proportion is due to the temporal change from the reference units of g/dL to g/L for UK laboratories in 2012. A transformation was then applied to approximately 30.87% of the values before use (i.e. these were values over 26.5 which would need to be divided by ten to be in the same units 74,769/242,214). Approximately 0.0045% of results (11 results out of 242,214 observations were outside the 1.6-26.5 range) were excluded as they were unlikely to contain valid Hb results.

The output is a series of rules which will identify valid Hb results for all patients in the THIN database. Further work may be required to exclude individual influential outlying values from specific analyses.

Data Source

This analysis used a 1% random sample of all patients from THIN1601 as its data source.

Analysis Tools

Stata version 14 was used to tabulate the frequency of Read code and value label combinations. The main commands used to describe the data were 'tabulate' and 'histogram'.

Procedure

1. Identification of ahdcode(s) likely to contain Hb values.
2. Identification of a suitable reference distribution for Hb.
3. Tabulation of Read code and value label combinations by frequency.
4. Review of individual candidate distributions.
5. Identification of GP requested plausible minimum and maximum values to be applied.
6. Description of the final method: a set of rules to be applied to identify valid Hb values.

4.4 Results: Development of a method to identify haemoglobin concentration values in THIN

4.4.1 Identification of ahdcode(s) likely to contain Hb values

The system lookup ahdcodes.dta were used to identify codes of interest to search in the ahd file (ahd.dta).

The following codes were identified:

datafile	ahdcode	description
TEST	1001400317	Haematology screening tests
TEST	1001400277	Carboxyhaemoglobin
TEST	1001400214	Haemoglobin variants
TEST	1001400044	Mean corpuscular haemoglobin
TEST	1001400027	Haemoglobin

Table 18: Potential ahdcodes which could contain Hb values.

The 1001400027 'Haemoglobin' ahdcode was the most relevant code to be used for data extraction.

The mean corpuscular haemoglobin of a sample is the average mass of haemoglobin in each red blood cell, and carboxyhaemoglobin measures haemoglobin combined with carbon monoxide.[19] These codes were therefore not included for further investigation.

In addition, 'Haematology screening tests' when investigated in the ahd table did not give medcodes relating to haemoglobin concentration. See **Table 19** below.

description	Frequency	Percent	Cumulative %
APTT inhibitor screening test	1	0.02	0.02
Haematology screening test	17	0.27	0.29
Haemoglobinopathy screening test	585	9.27	9.56
Haemolysis screening test	5,517	87.45	97.00
Sickle cell disease screening test	1	0.02	97.02
Sickle cell test negative	15	0.24	97.26
Sickle solubility test	129	2.04	99.30
Thrombophilia screening test	44	0.70	100.00
Total	6,309	100	

Table 19: Medcode descriptions associated with the ahdcode relating to 'Haematology Screening Tests'.

Finally, 'Haemoglobin variants' also did not give medcodes relating to the overall haemoglobin concentration and so this ahdcode was excluded from further analysis (**Table 20**).

description	Frequency	Percent	Cumulative %
Electrophoresis - Hb	254	4.06	4.06
Haemoglobin A	369	5.90	9.97
Haemoglobin A2 level	1,275	20.40	30.37
Haemoglobin C level	11	0.18	30.54
Haemoglobin D level	11	0.18	30.72
Haemoglobin E level	3	0.05	30.77
Haemoglobin F	1,495	23.92	54.69
Haemoglobin O level	1	0.02	54.70
Haemoglobin S Level	172	2.75	57.46
Haemoglobin acid electrophoresis	3	0.05	57.50
Haemoglobin alkaline electrophoresis	133	2.13	59.63
Haemoglobin electrophoresis	1,520	24.32	83.95
Haemoglobin variant NOS	65	1.04	84.99
Haemoglobin variant test	392	6.27	91.26
Haemoglobin variants	106	1.70	92.96
Haemoglobinopathy DNA studies	10	0.16	93.12
Kleihauer test	4	0.06	93.18
Methaemoglobin level	8	0.13	93.31
Oxyhaemoglobin level	3	0.05	93.36
Red cell Haemoglobin A2 estimation	391	6.26	99.62
Red cell haemoglobin S estimation	22	0.35	99.97
Unstable haemoglobin level	2	0.03	100.00
Total	6,250	100.00	

Table 20: Medcode descriptions associated with the ahdcode relating to Haemoglobin Variants.

4.4.2 Identification of a suitable reference distribution for Hb

Reference values for an adult can be obtained from haematology textbooks, path labs, haematologists, or from internet sources (e.g. http://labtestsonline.org.uk/understanding/analytes/haemoglobin/tab/test)

The Oxford Handbook of Clinical Haematology[20] gives the following reference values for an adult:

Normal values in an adult are approximately:

-130 to 180 g/L (13 to 18 g/dL) of blood for males

-115-165g/L (11.5-16.5 g/dL) of blood for females

In most cases, the most common Read code and unit value label for the ahdcode of interest (identified from the ahd.dta file) is likely to be the lab result of interest, reported in the units of interest. This combination was taken as the *reference distribution* as it included the normal ranges described above. A further visual comparison of this Hb distribution with the distribution from an external data source was used to validate this decision.

4.4.3 Tabulation of Read code and value label combinations by frequency

The field called data2 in the ahd table contained continuous numeric values for records with the ahdcode of interest. For Read code (medcode field) and value label (data3 field) combinations, values were dropped if they were equal to 0, as this is biologically implausible and if they were missing.

A new variable was generated called ‘value’ by converting the values stored as text in data2 into numeric format so that the distribution could be observed.

The medcodes were then analysed to see which ones were relevant for Hb continuous values (**Table 21**).

Medcode	description	Frequency
423..00	Haemoglobin estimation	228570
423..11	Hb estimation	12922
4237.00	Haemoglobin normal	330
424..00	Full blood count - FBC	135
4235.00	Haemoglobin low	126
423Z.00	Haemoglobin estimation NOS	58
4236.00	Haemoglobin borderline low	31
4239.00	Haemoglobin high	15
4234.00	Haemoglobin very low	9
42J..00	Neutrophil count	9
423B.00	Haemoglobin abnormal	8
4238.00	Haemoglobin borderline high	3
4232.00	Haemoglobin requested	2
D21z.00	Anaemia unspecified	2
4231.00	Haemoglobin not estimated	1
4233.00	Haemoglobin - sample sent	1
423C.00	Haemoglobin H inclusion	1

Table 21: All medcodes recorded under the ahd code for haemoglobin (1001400027).

The 42J..00 (neutrophil count) medcode was removed as this is not a relevant medcode for Hb. The other medcodes were included as they relate to Hb concentration.

The data3 column gives the unit 'value labels' for this subset of data. The text description (lookupdesc field) for each value label (data3 field) was derived from the system lookup table called ahdlookups.dta (**Table 22**).

data3	lookupdesc	Frequency
MEA056	g/dL	162903
MEA057	g/L	66320
MEA000	null value	6885
		5717
<None>		296
g/dl.		32
g/L		27
MEA001	%	13
g/l		11
g/dL.		5
MEA194	g/mol	3
MEA049	g	2
MEA015	/day	1
MEA026	1	1
MEA037	10*9/L	1
MEA051	g(hgb)	1
MEA055	g/d	1
MEA080	mg	1
MEA097	mmol/mol	1
MEA114	pg	1
gms/dl		1

Table 22: Unit value labels recorded under data3

The most frequent unit was g/dL, as this was the most commonly used unit in UK laboratories. There were also high numbers of values using the unit g/L due to laboratories standardising their units from 2012 onwards.[18] The g/L unit would require conversion to match the reference distribution in this circumstance. This was taken into account when looking at the different value label and medcode combinations.

All the possible medcode and unit code combinations were generated and ordered by most frequent. The first group combinations were labelled covering a cumulative percentage of over 99% of the data (**Table 23**).

Group Combination number	data3	medcode	Frequency	Percentage of Total	Cumulative %
1	MEA056	423..00	153671	0.634	0.634
2	MEA057	423..00	64853	0.268	0.902
3	MEA056	423..11	8779	0.036	0.938
4	MEA000	423..00	6496	0.027	0.965
5		423..00	3185	0.013	0.978
6		423..11	2512	0.010	0.989
7	MEA057	423..11	1446	0.006	0.995

Table 23: Unit value labels and medcode combinations. Blank boxes relate to when there is no value label. assigned.

Median values were generated for each Read code (medcode field) and value label (data3 field) combination, this helps when there are bimodal/multimodal distributions (**Table 24**). For example, the unit/medcode combination for different peaks in the distribution can be identified, which allows separation/conversion of the data if required.

data3	medcode	Median value of Hb	Frequency
MEA056	423..00	13.40	153671
MEA057	423..00	135.00	64853
MEA056	423..11	13.20	8779
MEA000	423..00	13.70	6496
	423..00	108.00	3185
	423..11	131.00	2512
MEA057	423..11	132.00	1446
<None>	423..00	13.50	296
MEA056	4237.00	13.50	172
MEA000	423..11	13.55	154
MEA000	4237.00	12.90	152
MEA056	424..00	13.55	134
MEA056	4235.00	10.90	66
MEA056	423Z.00	13.50	53
MEA000	4235.00	11.80	38
g/dl.	423..00	13.20	28
MEA000	4236.00	11.30	27
g/L	423..11	134.00	19
	4235.00	11.40	15

g/l	423..00	142.00	11
MEA056	4239.00	120.00	11
g/L	423..00	121.00	8
MEA057	4235.00	100.00	7
MEA000	423B.00	10.60	7
MEA001	423..00	8.00	7
MEA056	42J..00	4.57	7
MEA000	4234.00	82.00	6
g/dL.	423..00	13.40	5
MEA057	4237.00	125.00	5
MEA057	423Z.00	134.00	4
g/dl.	423..11	10.75	4
MEA001	423..11	3.35	4
MEA000	4239.00	133.00	4
MEA057	4236.00	114.00	3
MEA194	423..00	132.00	3
MEA056	4238.00	15.30	2
MEA056	4232.00	11.40	2
MEA056	4234.00	5.60	2
	D21z.00	57.40	2
MEA001	42J..00	47.95	2
MEA049	423..00	12.45	2
MEA114	423..00	29.70	1
MEA057	4234.00	76.00	1
MEA080	423..00	115.00	1
MEA056	423B.00	11.50	1
MEA015	423..11	24.08	1
	4236.00	11.50	1
MEA097	423..11	54.00	1
gms/dl	423..00	12.00	1
MEA056	423C.00	14.10	1
	4237.00	12.30	1
MEA026	423..11	13.50	1
MEA055	423..00	14.20	1
MEA056	4231.00	144.00	1
MEA057	424..00	122.00	1

MEA051	423..11	13.00	1
	423Z.00	14.10	1
MEA037	423..00	12.30	1
MEA000	4238.00	14.90	1
MEA056	4233.00	13.00	1

Table 24: Median value of Hb for each unit value label and medcode combination. Blank boxes relate to when there is no value label.

4.4.4 Review of individual candidate distributions.

A new variable called 'combogroup' was generated and assigned a group combination number for each Read code (medcode field) and value label (data3 field) pair investigated (See **Table 23** above).

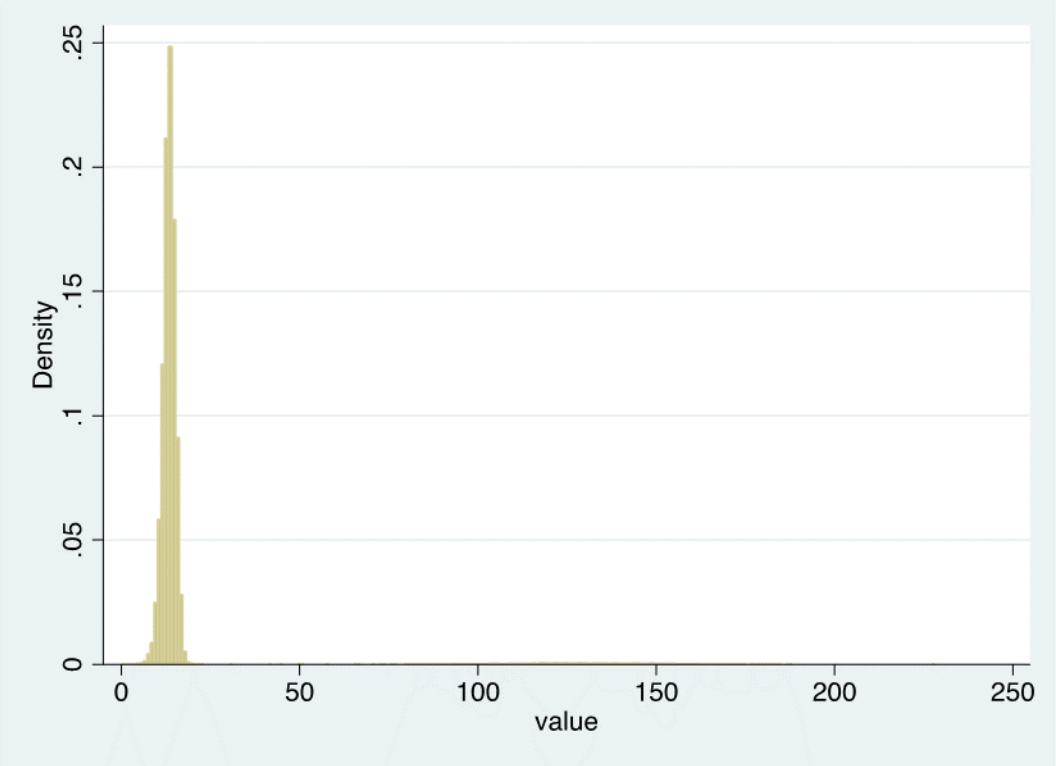
4.4.4.1 Reference Distribution Group Combination 1 (MEA056 and 423..00)

The reference distribution is the most common Read code (medcode field) and value label (data3 field) combination. The other most frequent combinations were compared to this group and the remaining small groups were then combined together for analysis once over 95% of the data was covered. In this case, the most frequent Read code and unit value label combination was when data3 was MEA056 (g/dL) and the medcode was 423..00 (haemoglobin estimation).

This pair was assigned as combogroup 1.

A histogram was produced to summarise the distribution and to determine the maximum and minimum values. Below is the code used in Stata to produce **Figure 10:**

```
hist value if medcode=="423..00" & data3=="MEA056", width(1)
summarize value if medcode=="423..00" & data3=="MEA056"
```



Variable	Obs	Mean	Std.Dev.	Min	Max
Value	153,671	15.4	15.8	0.2	228.0

Figure 10: The plot shows frequency density by haemoglobin value concentration for the reference distribution (MEA056 g/dL and medcode 423..00 haemoglobin concentration)

The reference range is between 11.5-18 g/dL of blood.

Based on the maximum value observed, there could be a distribution higher than expected for this medcode and value label pair. Therefore, the values of haemoglobin over 50 g/dL were investigated and all values for this pair were tabulated to analyse the frequencies at higher values. Below is the code used in Stata to produce **Figure 11.**

```
hist value if medcode=="423..00" & data3=="MEA056" & value>50, width(1) frequency
```

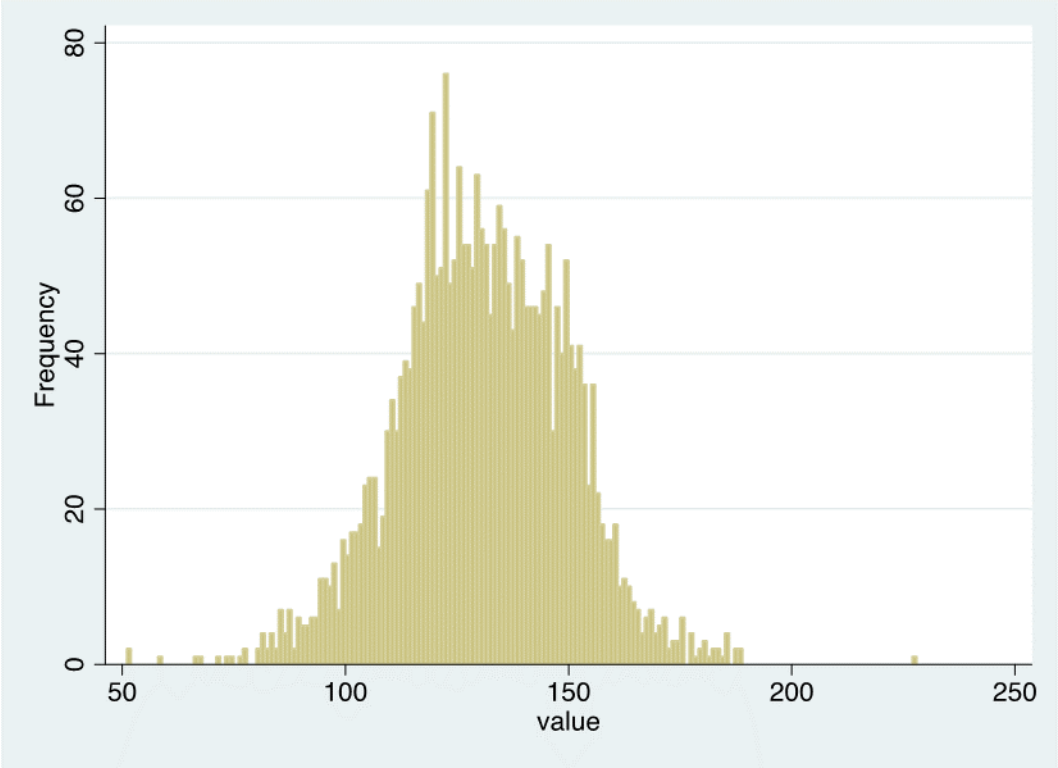


Figure 11: Hb values above 50 for the reference distribution (MEA056 g/dL and medcode 423..00 haemoglobin concentration).

This distribution is about ten times the value of the lower distribution; therefore it can be concluded that these values might have been given the incorrect value label. The next most common value label is 'g/L' which is ten times the 'g/dL' unit.

Based on the GP requested plausible values from the external data source a cut-point was chosen to separate these two distributions.

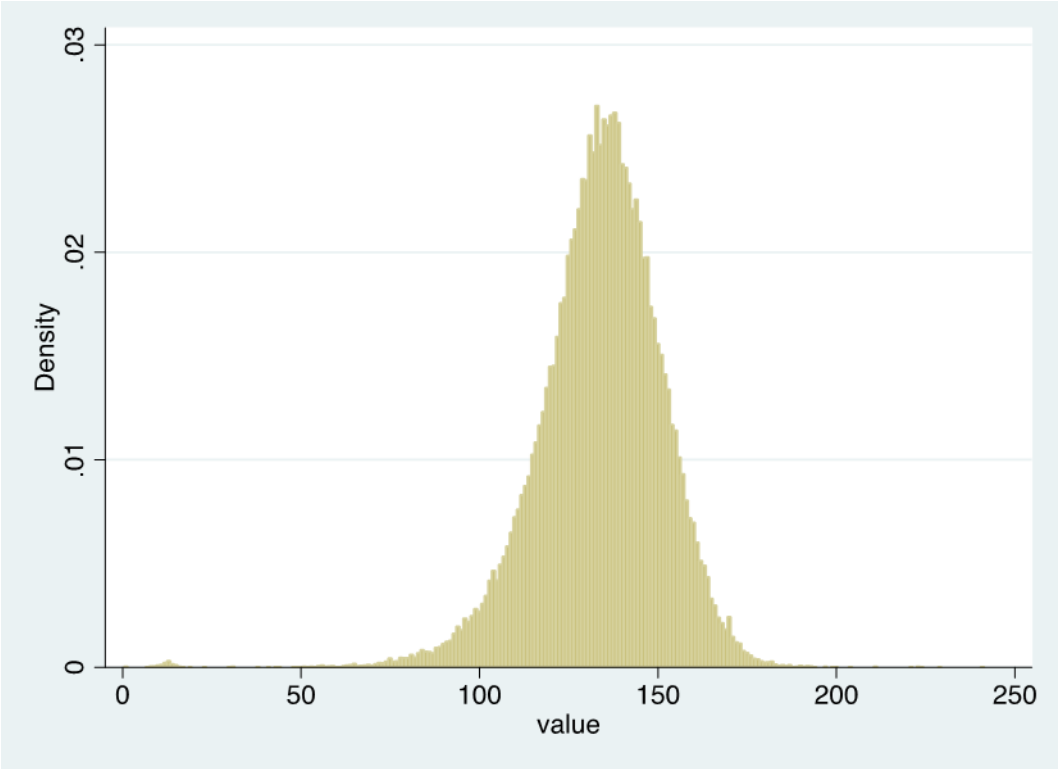
Conclusion: If the value is over 26.5 for this unit then divide by 10 based on the GP requested plausible ranges.

4.4.4.2 Group Combination 2 (MEA057 and 423..00)

The next most common medcode and value label combination was compared with the reference distribution. This was MEA057 g/L and 423..00 haemoglobin estimation. This combination was assigned as combogroup 2.

The code used in Stata to produce the histogram and summary statistics for **Figure 12** is shown below.

```
hist value if medcode=="423..00" & data3=="MEA057", width(1)
summarize value if medcode=="423..00" & data3=="MEA057"
```



Variable	Obs	Mean	Std.Dev.	Min	Max
Value	64,853	133.9	17.3	0.4	241.0

Figure 12: Distribution (Hb value) for MEA057 and 423..00.

Figure 12 shows that there could be a second lower peak at about 15 g/L in the distribution of values which is lower than expected for this medcode and unit pair. Therefore, the values of Hb lower than 50 g/L were tabulated for this pair to see the frequencies/any peaks at lower values (**Figure 13**). See Stata code below:

```
hist value if medcode=="423..00" & data3=="MEA057"& value<50, frequency
```

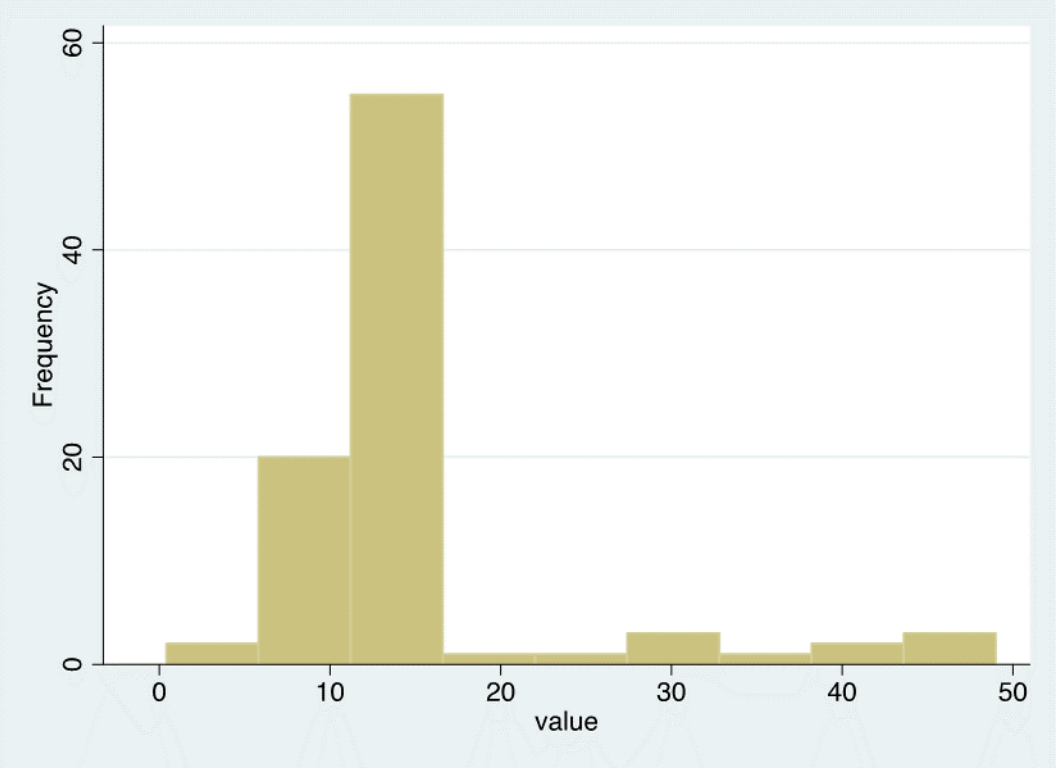


Figure 13: Hb values below 50 for combination group 2 (MEA057 and 423..00).

Based on the GP requested plausible values, a cut-point was chosen for these two distributions to utilise all the data.

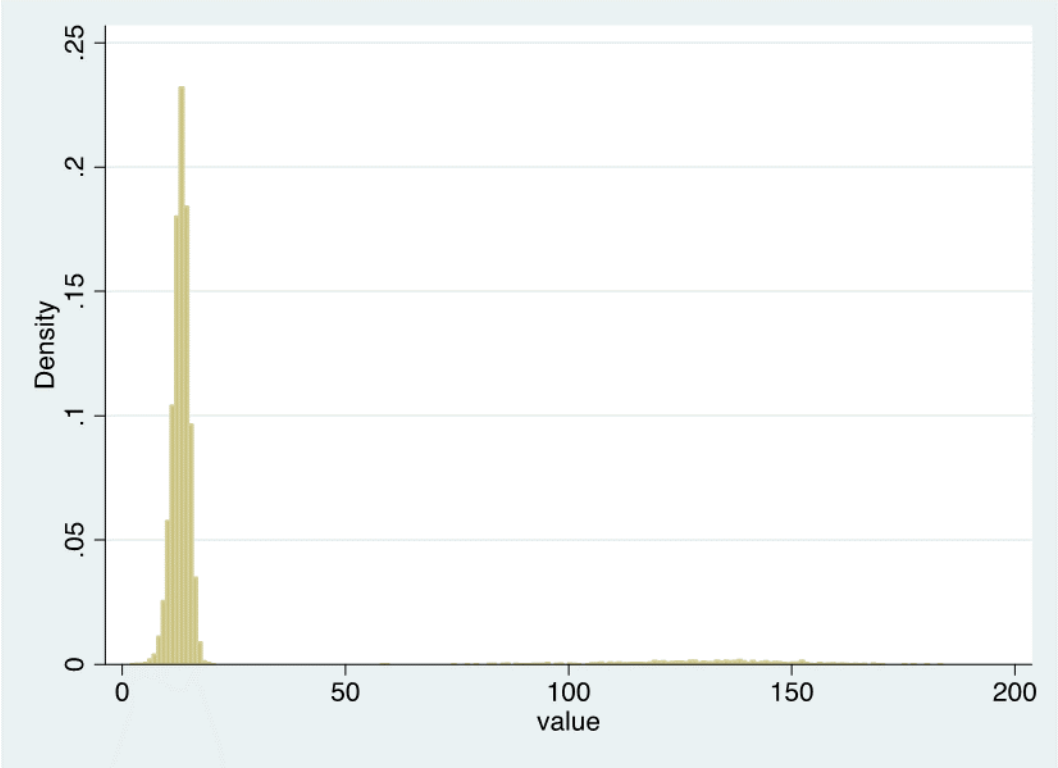
Conclusion: If the value is over 26.5 then divide by 10.

4.4.4.3 Group Combination 3 (MEA056 and 423..11)

The next most common value label and medcode pair was MEA056 g/dL and 423..11 (Hb estimation). This was assigned as combogroup 3.

Below is the Stata code used to produce **Figure 14**:

```
hist value if medcode=="423..11" & data3=="MEA056", width(1)
summarize value if medcode=="423..11" & data3=="MEA056"
```



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	8,779	19.5	27.2	1.8	183.0

Figure 14: Distribution (Hb value) for MEA056 and 423..11.

There is a potential higher distribution present since the reference range is between 11.5-18 g/dL of blood. Based on the maximum value, there could be a distribution higher than expected for this medcode and unit pair. Therefore, values of Hb over 50g/dL were tabulated for this pair to inspect the frequencies at higher values.

```
hist value if medcode=="423..11" & data3=="MEA056" & value>50, width(1)
```

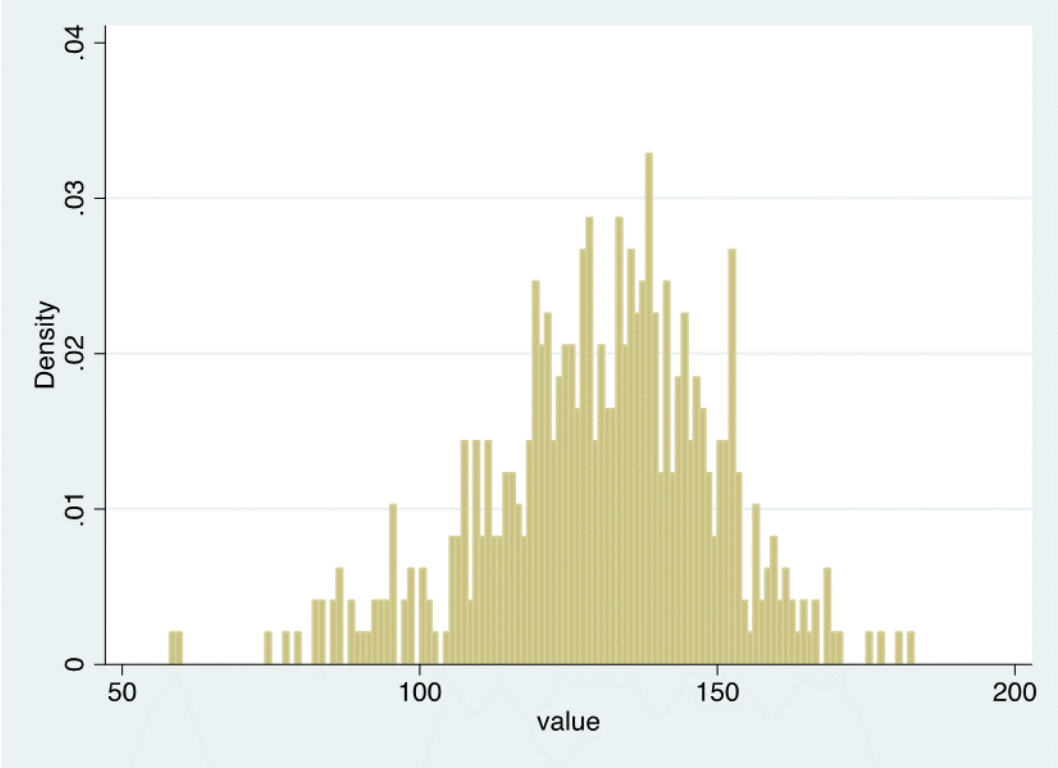


Figure 15: Hb values above 50 for combination group 3 (MEA056 and 423..11).

Again this is about ten times the value of the lower distribution, it can therefore be concluded that these values may have been given the incorrect unit label.

The next most common unit label is g/L which is ten times 12-18 g/dL of blood.

Based on the GP requested plausible values, a cut-point was chosen for these two distributions to utilise all the data.

Conclusion: If the value is over 26.5 then divide by 10.

4.4.4.4 Group Combination 4 (MEA000 and 423..00)

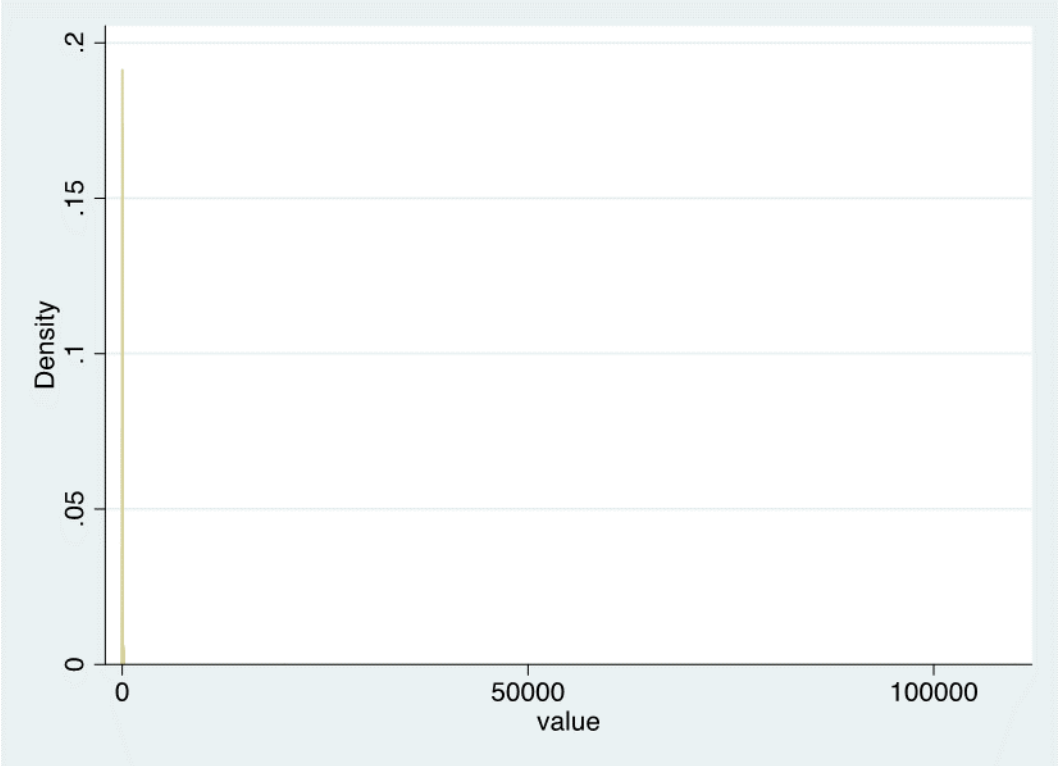
The next most common value label and medcode combination was MEA000 (null value) and 423..00 (haemoglobin estimation).

This was assigned as combogroup 4.

The histogram and summary statistics for **Figure 16** was produced using the following Stata code.

```
hist value if medcode=="423..00" & data3=="MEA000", width(1)

summarize value if medcode=="423..00" & data3=="MEA000"
```



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	6,496	53.9	1387.4	1.7	110000.0

Figure 16: Distribution (Hb value) for MEA000 and 423..00

Since there was a very large value (over 100,000) present for this combination, the distribution could not be visualized in detail. The data was therefore summarised and this identified that there was a maximum value of 110,000. This could not be explained by any other unit and was outside the GP requested plausible ranges.

To determine how many values there were in this top range, the data was tabulated using the following Stata code:

```
tabulate value if medcode=="423..00" & data3=="MEA000"
```

There were only one or two values this high (See **Table 25**) so it was concluded that these are potential outliers and have been inputted/recorded incorrectly.

Value	Frequency	Percent	Cumulative Percent
179	1	0.02	99.94
187	1	0.02	99.95
195	1	0.02	99.97
20000	1	0.02	99.98
110000	1	0.02	100.00

Table 25: Frequency of higher Hb values for MEA000 and 423..00.

Large values such as these will be taken into account when setting the GP requested plausible maximum value (e.g. exclude values over 265 g/L or 26.5 g/dL).

By limiting the Hb value to below 200, the distribution could be analysed more closely. The Stata code for **Figure 17** is given below:

```
hist value if medcode=="423..00" & data3=="MEA000" & value<200, width(1) frequency
```

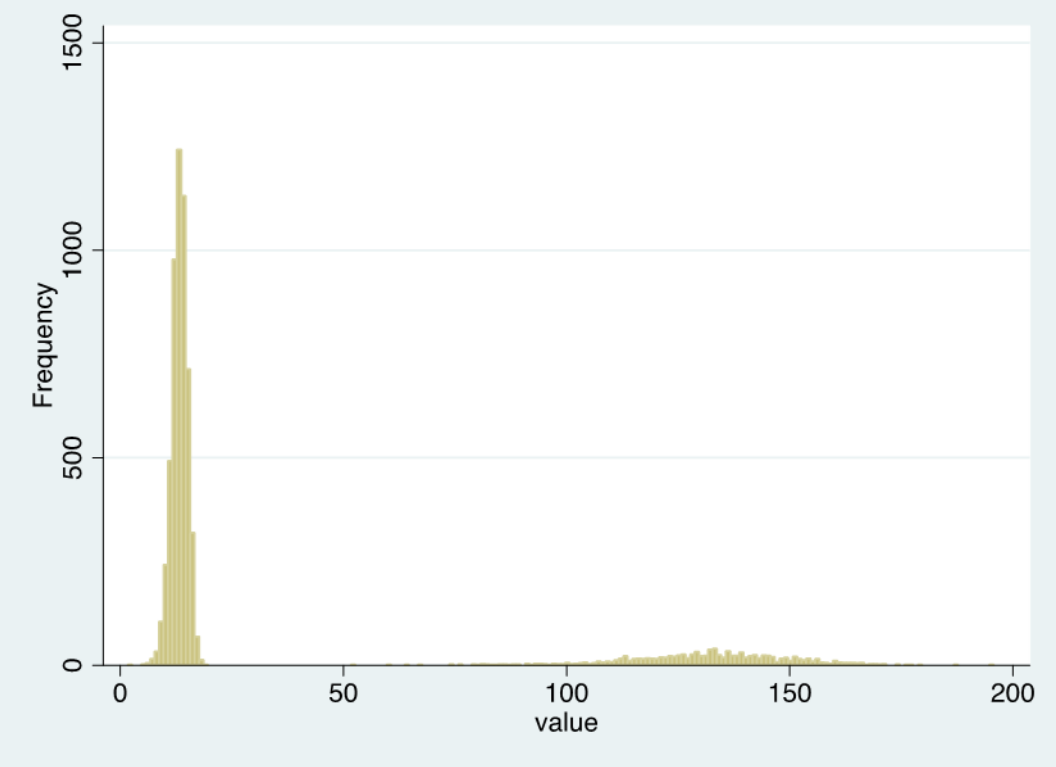


Figure 17: Hb values below 200 for combination group 4 (MEA000 and 423..00).

Two distributions were visible again as seen in the previous combination groups.

The cut-off was then set to above 50 but below 200, so the higher distribution could be visualised (**Figure 18**).

Stata code given below:

```
hist value if medcode=="423..00" & data3=="MEA000" & value>50 & value<200
```

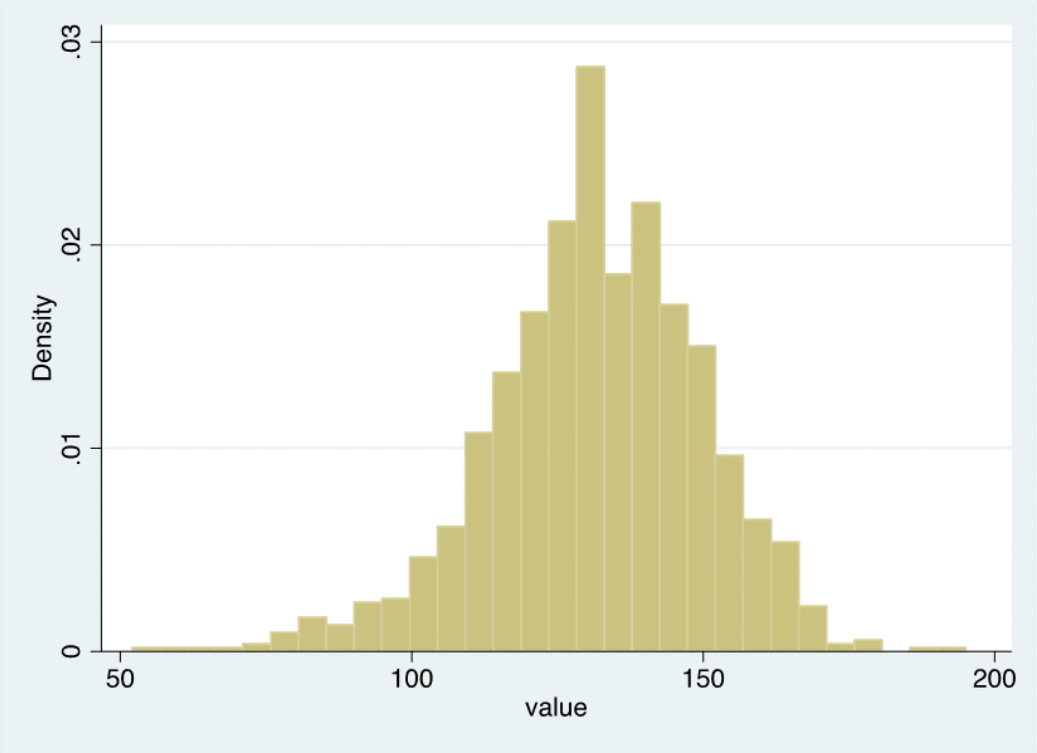


Figure 18: Hb values below 200 and over 50 for combination group 4 (MEA000 and 423..00).

The lower distribution was visualised by restricting the Hb value to below 50. Stata code given below (**Figure 19**):

```
hist value if medcode=="423..00" & data3=="MEA000" & value<50
```

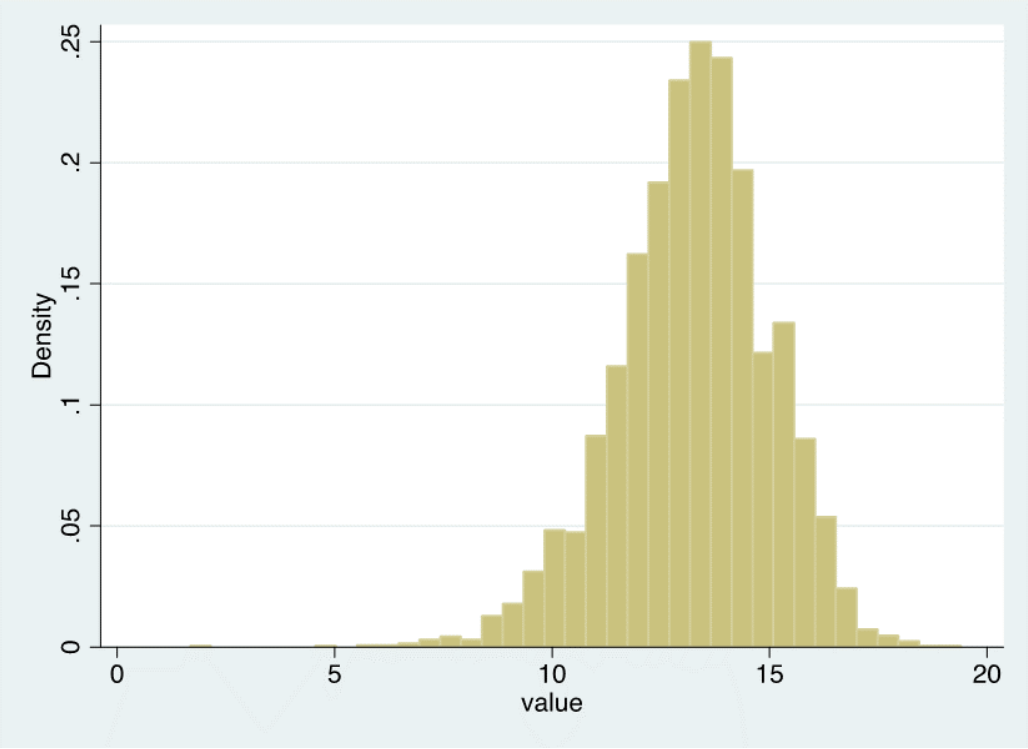


Figure 19: Hb values below 50 for combination group 4 (MEA000 and 423..00).

Based on the GP requested plausible values, a cut point was chosen for these two distributions to utilise all the data.

Conclusions: If the value is over 26.5 then divide by 10.

If the values are over 26.5 after conversion they can be excluded from the analysis as this is outside the GP requested range seen in the external data.

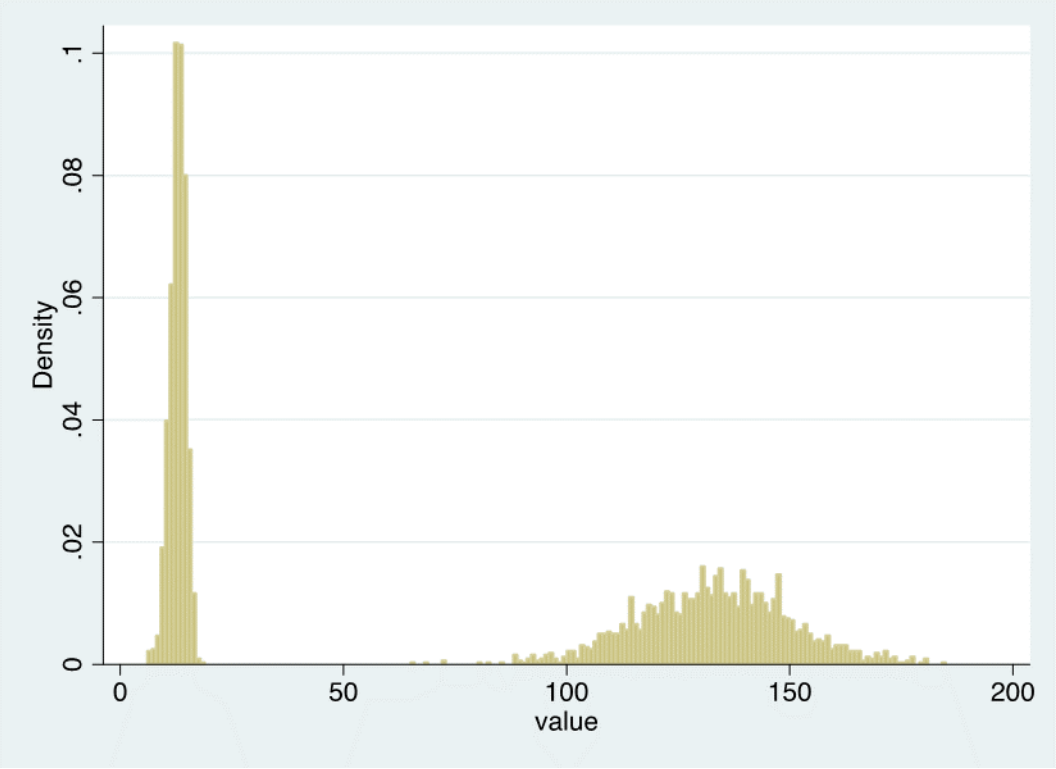
4.4.4.5 Group Combination 5 ('No unit' and 423..00)

The next most common combination was 'no unit' and 423..00.

This was assigned as combo group 5.

The histogram and summary statistics for **Figure 20** were produced using the following Stata code.

```
hist value if medcode=="423..00" & data3=="", width(1)
summarize value if medcode=="423..00" & data3==""
```



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	3,185	77.2	61.0	6.0	185.0

Figure 20: Distribution (Hb value) for 'No unit' and 423..00

Based on the GP requested plausible values a cut-point was chosen for these two distributions to utilise all the data.

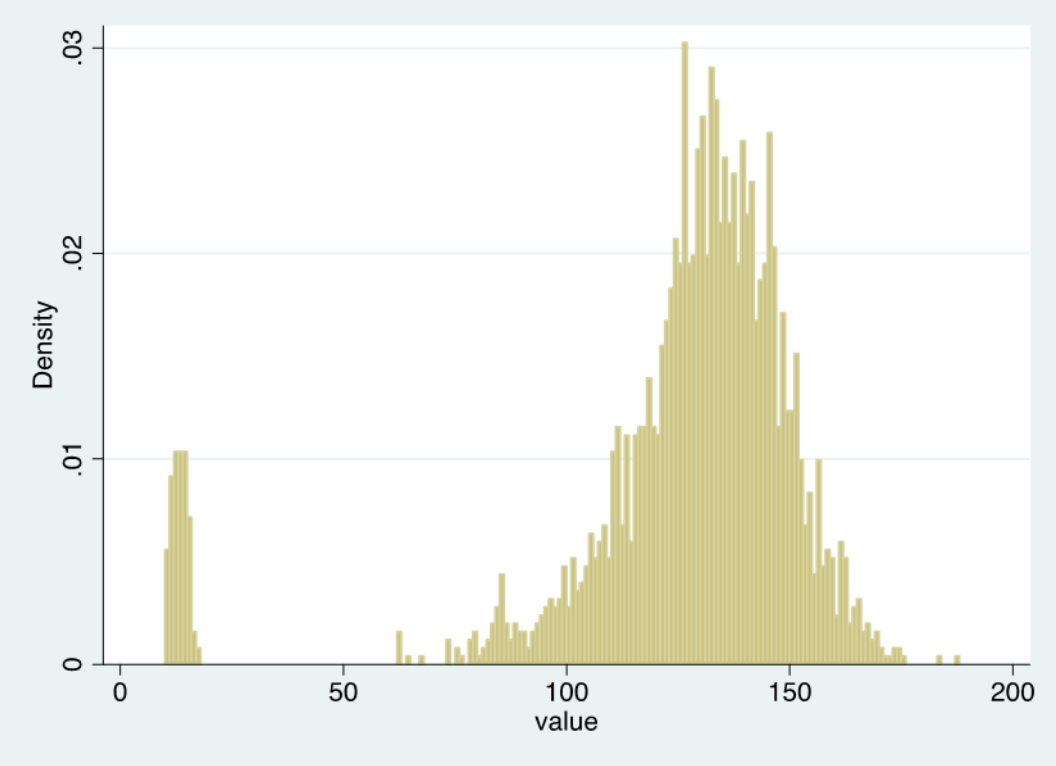
Conclusion: If the value is over 26.5 then divide by 10.

4.4.4.6 Group Combination 6 ('No unit' and 423..11)

The next most common combination was 'no unit' and 423..11. This was assigned as combo group 6 and the histogram and summary statistics for **Figure 21** were produced using the following Stata code:

```
hist value if medcode=="423..11" & data3=="", width(1)

summarize value if medcode=="423..11" & data3==""
```



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	2,512	124.4	31.8	10.0	188.0

Figure 21: Distribution (Hb value) for 'No unit' and 423..11

Based on the GP requested plausible values, a cut-point was chosen for these two distributions to utilise all the data.

Conclusion: If the value is over 26.5 then divide by 10.

4.4.4.7 Group Combination 7 (MEA057 and 423..11)

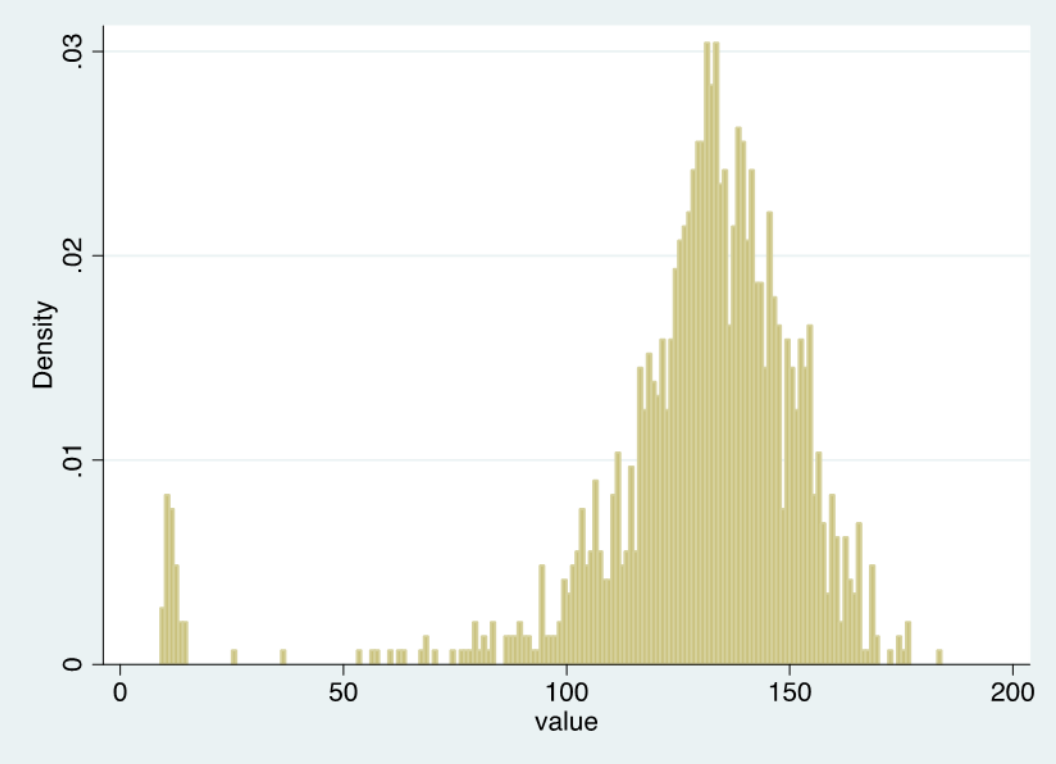
The next most common combination was MEA057 and 423..11.

This was assigned as combogroup 7.

The histogram and summary statistics for **Figure 22** were produced using the following Stata code:

```
hist value if medcode=="423..11" & data3=="MEA057", width(1)

summarize value if medcode=="423..11" & data3=="MEA057"
```



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	1,446	128.6	26.8	9.0	184.0

Figure 22: Distribution (Hb value) for MEA057 and 423..11

Based on the GP requested plausible values, a cut-point was chosen for these two distributions to utilise all the data.

Conclusion: If the value is over 26.5 then divide by 10.

4.4.4.8 Remaining Group Combinations

For the final step, the remaining infrequent value label and medcode combinations were investigated altogether. Ninety-nine per cent of the data have been covered using the combinations above so this was investigating a much smaller amount of data (**Figure 26**).

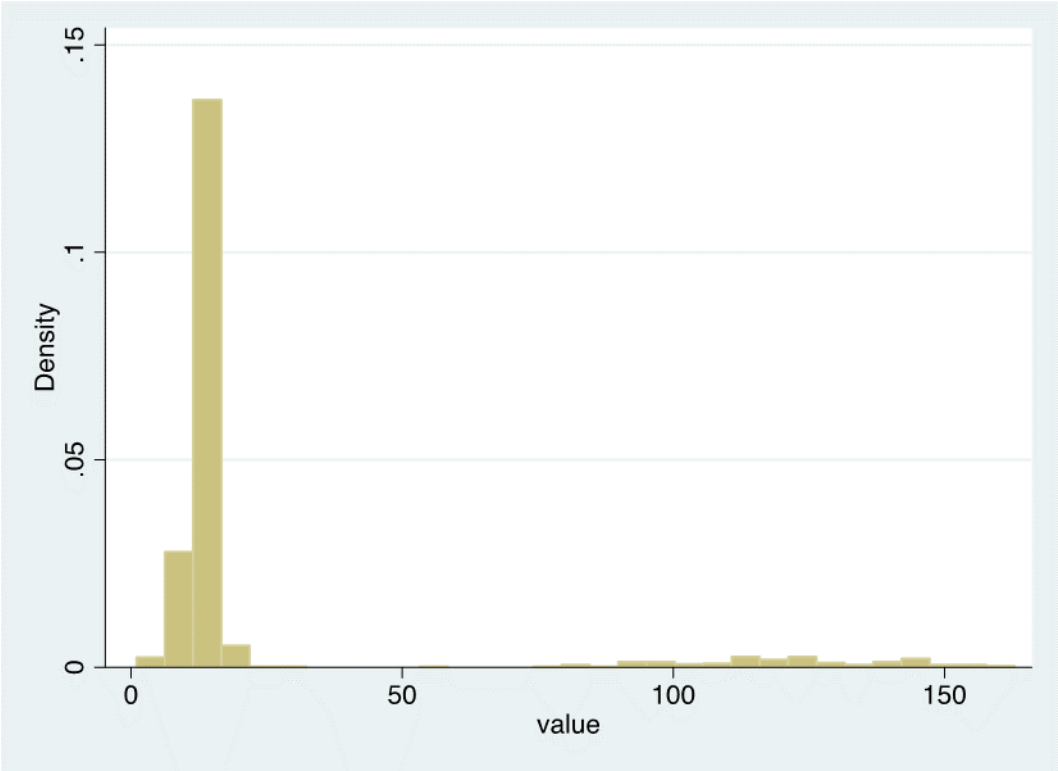
combogroup	Frequency	Percent	Cum.
1	153,671	63.44	63.44
2	64,853	26.78	90.22
3	8,779	3.62	93.84
4	6,496	2.68	96.53
5	3,185	1.31	97.84
6	2,512	1.04	98.88
7	1,446	0.60	99.47
(all remaining combinations) .	1,272	0.53	100.00
Total	**242,214**	**100**	

Table 26: Frequency of different group combinations of medcodes and value labels.

The following Stata code was used to summarise the remaining group combinations:

```
hist value if combogroup==.,
summarize value if combogroup==.
```

The distribution of the remaining group combinations is shown in **Figure 23**.



Variable	Obs	Mean	Std. Dev.	Min	Max
Value	1,272	23.6	32.6	1.0	163.0

Figure 23: Distribution (Hb value) for all remaining group combinations.

The histogram shows a bimodal distribution as observed in previous combinations.

Hb values below 50 were then investigated using the following Stata code (**Figure 24**):

```
hist value if combogroup==. & value<50, frequency
```

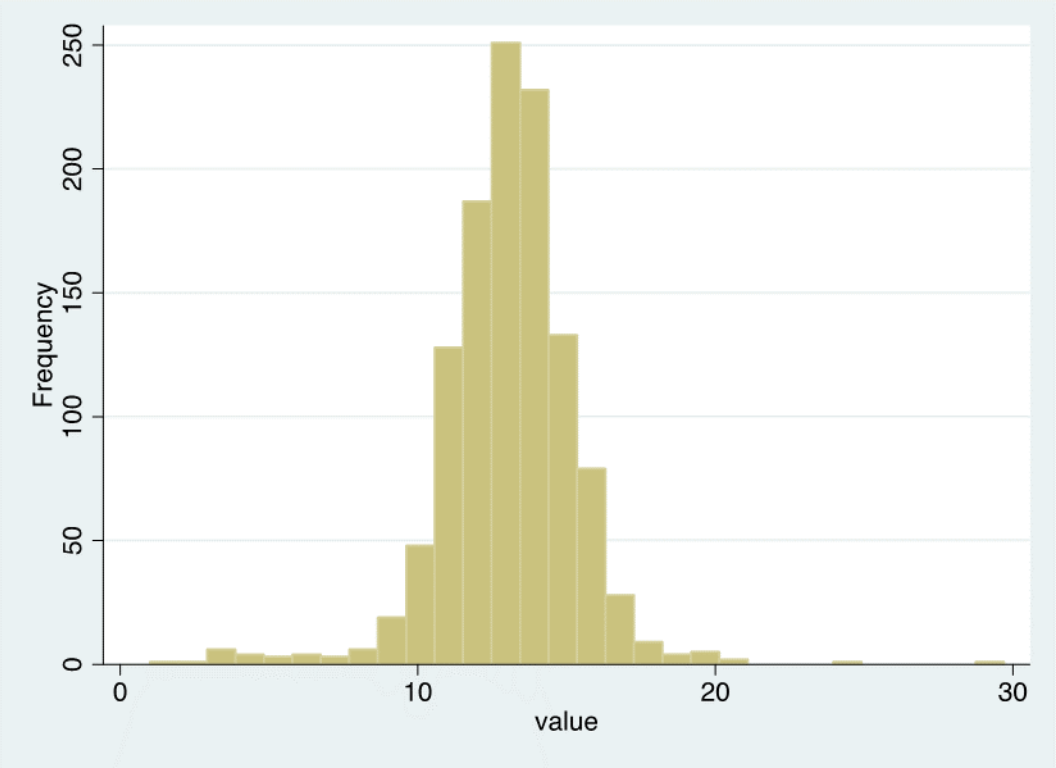


Figure 24: Hb values below 50 for remaining group combinations.

To investigate the distribution for values over 50 the following Stata code was used (**Figure 25**):

```
hist value if combogroup==. & value>50, frequency
```

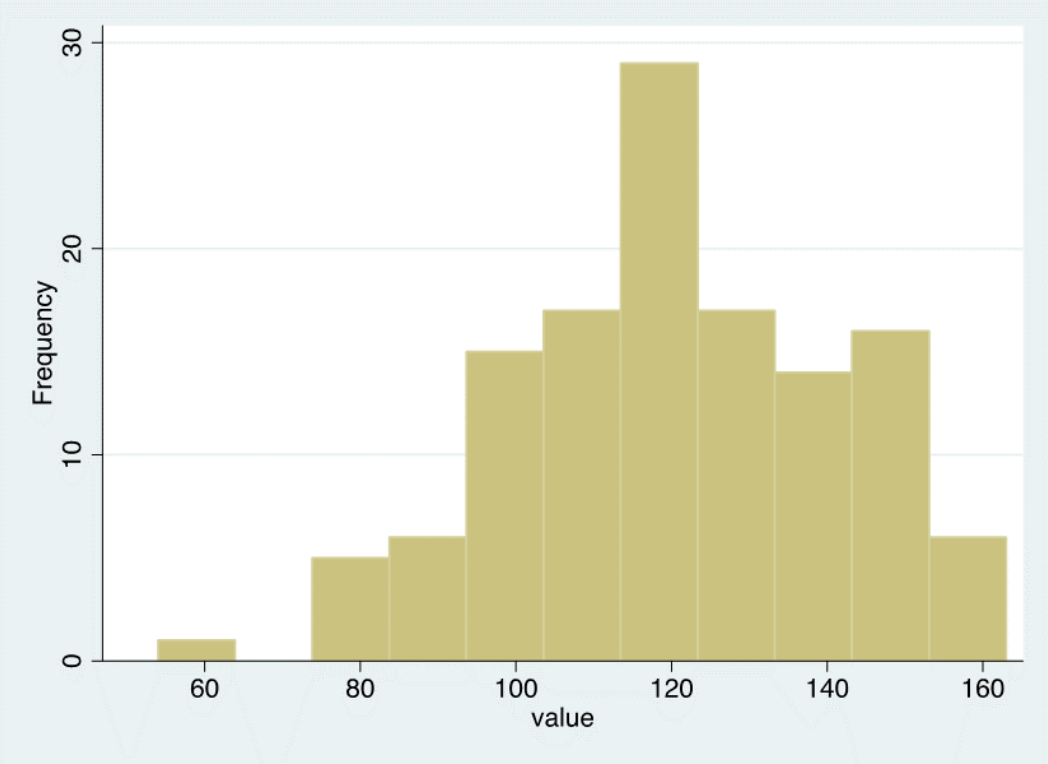


Figure 25: Hb values above 50 for remaining group combinations

Based on the GP requested plausible values, a cut-point was chosen for these two distributions to utilise all the data.

Conclusion: If the value is over 26.5 then divide by 10.

4.4.5 Identification of GP requested plausible minimum and maximum values to be applied

The distribution of values obtained from a local hospital laboratory for GP requested Hb tests in the past three years was compared with the observed THIN distribution.

The distribution of Hb values from local lab data is shown in **Figure 26**. This distribution is for both males and females, for all ages in the recommended g/L units. The minimum result reported by the laboratory was 16 g/L and the maximum was 265 g/L. Summary statistics for this local lab data are also shown in **Table 27**.

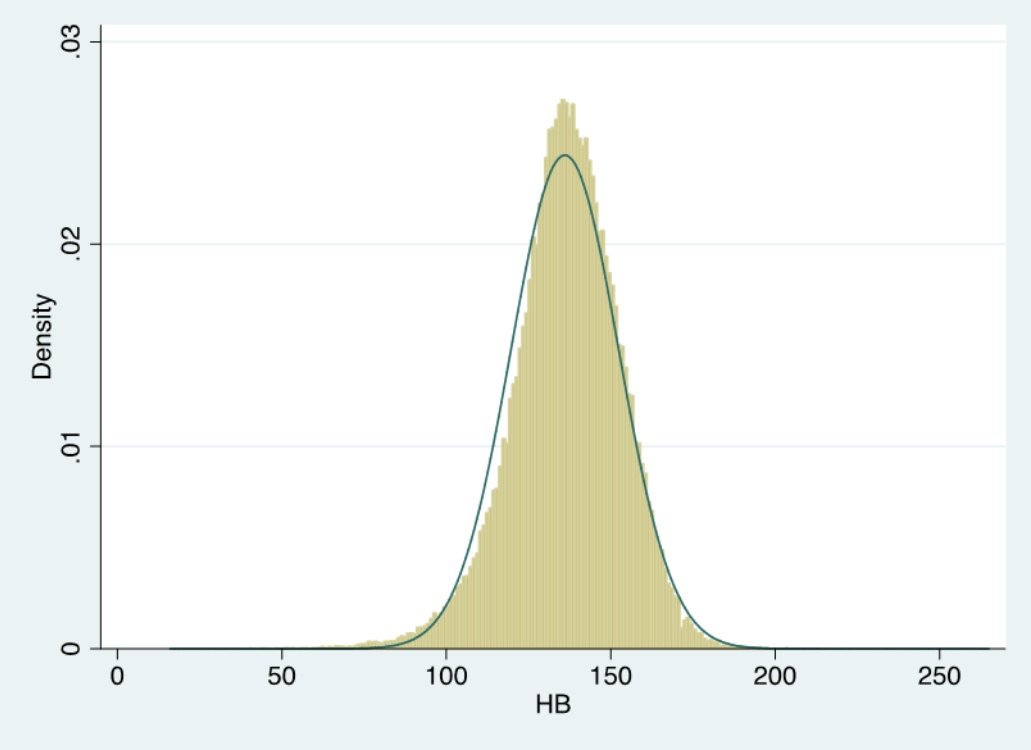


Figure 26: Distribution of Hb values from local lab data for GP requested tests.

	Percentiles	Smallest		
1%	91	16		
5%	108	32		
10%	116	40	**Obs**	103,801
25%	126	40	**Sum of Wgt.**	103,801
50%	137		**Mean**	136.04
		Largest	**Std. Dev.**	16.35
75%	147	246		
90%	156	252	**Variance**	267.39
95%	161	256	**Skewness**	-0.39
99%	172	265	**Kurtosis**	4.13

Percentile	Hb Value
0.001	32
0.01	48
0.1	68
1	91
25	126
50	137
75	147
99	172
99.9	185
99.99	213
99.999	256

Table 27: Summary statistics for Hb values from local lab data for GP requested tests

For the reference distribution derived from THIN (the most common medcode and unit value label combination), the following values/distribution are shown in **Table 28** and **Figure 27**. Since there was a bimodal distribution for the reference distribution due to the presence of two different units, the Hb value was restricted to below (or equal to) 26.5 for these investigations. The following Stata code is listed below.

```
summarize value if medcode=="423..00" & data3=="MEA056" & value<=26.5, detail

hist value if medcode=="423..00" & data3=="MEA056" & value<=26.5


_pctile value  if medcode=="423..00" & data3=="MEA056" & value<=26.5, p(0.001, 0.01,
0.1, 1, 25, 50, 75, 99, 99.9, 99.99, 99.999)


return list
```

	Percentiles	Smallest		
1%	8.7	0.2		
5%	10.4	0.42		
10%	11.2	0.9	**Obs**	150,912
25%	12.3	0.95	**Sum of Wgt.**	150,912
50%	13.4		**Mean**	13.31
		Largest	**Std. Dev.**	1.69
75%	14.4	21.1		
90%	15.4	21.3	**Variance**	2.859
95%	15.9	22.4	**Skewness**	-0.449
99%	16.9	22.4	**Kurtosis**	3.99

Percentile	Hb Value
0.001	0.4
0.01	3.4
0.1	6.4
1	8.7
25	12.3
50	13.4
75	14.4
99	16.9
99.9	18.1
99.99	19.6
99.999	22.4

Table 28: Summary statistics for Hb values for the THIN reference distribution (limited to below 26.5 g/dL due to the bimodal distribution).

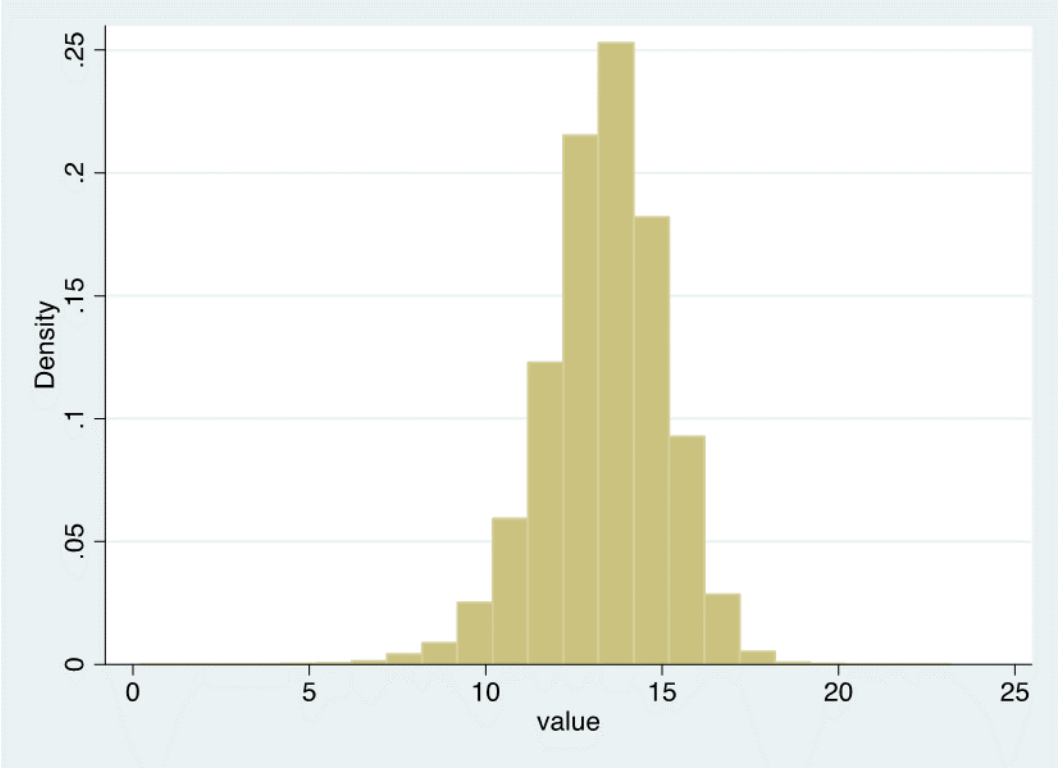


Figure 27: Distribution of Hb values from the THIN database for the reference distribution (restricted to below 26.5 g/dL due to the bimodal distribution).

By limiting the data to below 26.5 g/dL these THIN results roughly match the local lab data (when converting units by multiplying by ten) particularly when comparing median values and the density. Therefore, this external data can be used for defining minimum and maximum results to the derived haemoglobin concentrations.

Minimum – 16 g/L (1.6 g/dL)

Maximum – 265 g/L (26.5 g/dL)

4.4.6 Description of the final method: a set of rules to be applied to identify valid Hb values.

Based on all of the above investigations the following rules can be applied to extract data for patients with a continuous Hb concentration result.

1. Select all records with ahdcode 1001400027 (This is haemoglobin).

2. Drop records with medcode 42J..00. (This is the medcode for neutrophil count.)

3. If the value in the data2 field is over 26.5 then divide by ten (these are mostly g/L) to convert them to the reference units in this case g/dL.

e.g.

```
generate value = data2
replace value = (value/10) if value>26.5 & !missing(value)
```

4. Drop records where the value is outside the range 1.6-26.5. (These are the minimum and maximum values reported by the hospital laboratory for GP initiated Hb. This will also remove any values which are missing or equal to 0.)

5. Drop records where the event date is missing. (They cannot be used in any analyses without a date.)

5.0 Compiling Read Code Lists

In order to extract the relevant data in **Chapter 5**, a clinical code/Read code list was generated for over 30 different clinical features and risk factors. Translation of symptoms into clinical Read codes was informed by previous work by Professor Tom Marshall which analysed predictors of colorectal cancer.[21] In addition, the ClinicalCodes repository set up by the University of Manchester and funded by the NIHR was also used to help develop code lists (https://clinicalcodes.rss.mhs.man.ac.uk/).[22]

The methodology described by Davé and Peterson[3] along with the THIN Data Guide for Researchers[1] was used to help inform a strategy for code set engineering. An evolving and iterative search strategy was used to search the Read code dictionary provided with THIN for key words using the 'regexm' function implemented in Stata. From the keyword search it was then possible to identify relevant parent and child stems for Read codes. These additional stems were then browsed for inclusion and all possible terms relating to a particular symptom or/diagnosis were extracted into an Excel Spreadsheet. This resulting Read code list was then subjected to review by one individual before being second reviewed by a clinician with expertise within that field for code set validation. This final Read code list was then used to extract bowel cancer screening diagnoses from the THIN database for analysis.

In some instances, previous code lists were examined to identify keywords for the iterative search strategy and to examine Read code stems to ensure no key terms were missed.

5.1 Read Code List Development for Bowel Cancer Diagnosis

This section reports the derivation of a Read code list for Bowel Cancer Diagnosis; the key outcome used for analysis in **Chapter 5**. An evolving and iterative search strategy was used to search the Read Code dictionary for key words using the 'regexm' function (regular expressions) implemented in Stata. This search strategy was kept broad since detailed review of the preliminary list will remove any irrelevant terms. Key terms were discussed with another reviewer/clinician and previous lists from ClinicalCodes and other previous research were used when compiling the search.

Words encompassing the region of interest (bowel, colon, caecum, gastrointestinal tract etc) along with synonyms for cancer (e.g. neoplasia, carcinoma, cancer, adenocarcinoma) were combined in the search strategy. The final search strategy developed in Stata is shown in **Table 29**.

```
replace case=1 if regexm(lcase,"colo.*cancer")
replace case=1 if regexm(lcase,"colo") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"bowel") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"rect") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"caec") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"append") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"git") & regexm(lcase,"neop|carcinoma|adeno|cancer")

browse if case ==1
```

Table 29: Key word search strategy for bowel cancer diagnosis developed in Stata 14.

The Read code dictionary includes a description with the relevant Read code and associated stems. The resulting list of Read codes and the corresponding descriptions from the initial search were examined and saved in an Excel Spreadsheet. A note was made for any codes which could be removed based on this initial review. For instance, most of the bowel cancer screening codes were not relevant for this particular term. In addition, codes such as the 'QCancer colorectal cancer risk calculator' (38GT000) could be marked for exclusion. See **Table 30** for an extract of the key word search results in Stata.

	medcode	description
11664	38GT000	QCancer colorectal cancer risk calculator
22756	6864.00	Large bowel neoplasm screen
22757	6864.11	Colon neoplasm screen
22758	6864.12	Rectal neoplasm screen
22760	6866.00	Bowel cancer screening programme: faecal occult blood result
23209	68W2.00	Bowel cancer screening programme
23210	68W2000	Bowel Cancer Screening Programme bowel scope screening test
23214	68W2400	Bowel scope (flexible sigmoidoscopy) screen: cancer detected
23215	68W2500	Bowel scope (flexi-sig) screen: suspected cancer detected
41915	8CAo.00	Patient given advice about bowel cancer
43488	8Hn1.00	Fast track referral for suspected gynaecological cancer
43495	8Hn4.00	Fast track referral for suspected colorectal cancer
44050	8IA3.00	Bowel cancer screening declined
44402	8OA5.00	Prov of written info about bowel cancer screening programme
47935	9Ni2.00	Did not attend bowel cancer screening programme nurse clinic
47937	9Ni3.00	Did not attend bowel cancer screening
47956	9Nic000	DNA fast track suspected gynaecological cancer clinic
48360	9Np7.00	Seen in fast track suspected colorectal cancer clinic
49261	9Ow..00	Bowel cancer screening programme administration
49264	9Ow1.00	Bowel cancer detected by national screening programme
49266	9Ow2.00	No response to bowel cancer screening programme invitation
49269	9Ow3.00	Not eligible for bowel cancer screening programme
49270	9Ow4.00	Bowel cancer screening programme telephone invitation
49272	9Ow5.00	Bowel cancer screening programme invitation letter sent
50356	A4z1.00	Adenoviral meningitis
52200	B13..00	Malignant neoplasm of colon
52201	B130.00	Malignant neoplasm of hepatic flexure of colon
52202	B131.00	Malignant neoplasm of transverse colon
52203	B132.00	Malignant neoplasm of descending colon
52204	B133.00	Malignant neoplasm of sigmoid colon
52205	B134.00	Malignant neoplasm of caecum
52206	B134.11	Carcinoma of caecum
52207	B135.00	Malignant neoplasm of appendix
52208	B136.00	Malignant neoplasm of ascending colon
52209	B137.00	Malignant neoplasm of splenic flexure of colon
52210	B138.00	Malignant neoplasm, overlapping lesion of colon
52211	B139.00	Hereditary nonpolyposis colon cancer
52212	B13y.00	Malignant neoplasm of other specified sites of colon
52213	B13z.00	Malignant neoplasm of colon NOS

Table 30: Extract of results from Stata using the key word search complied for bowel cancer diagnosis

From this initial list, the Read code hierarchies of relevance could be examined. For bowel cancer diagnosis the code stems B….00 (Neoplasms), B….11 (Cancers) and 68…00 (Screening) were of relevance.

Since Read codes are hierarchical, all relevant parent codes along with the child codes (and sibling codes) were assessed for inclusion above and below these values (see **Table 31** for an example).

Read Code	Description	Code Type
B13..00	Malignant neoplasm of colon	Parent code
B130.00	Malignant neoplasm of hepatic flexure of colon	Child code
B131.00	Malignant neoplasm of transverse colon	Child code
B132.00	Malignant neoplasm of descending colon	Child code
B133.00	Malignant neoplasm of sigmoid colon	Child code
B134.00	Malignant neoplasm of caecum	Child code
B134.11	Carcinoma of caecum	Sibling code

Table 31: Hierarchical relationships between Read codes used for malignant neoplasm of Colon as an example.

By browsing the parent stems, the following relevant stems displayed in **Table 32** were of relevance. By looking at the hierarchies as well as the key word search, additional Read codes could be identified.

```
browse if regexm(medcode, "^B\..*|^B1.*|^B5.*|^B8.*|^B9.*|^BA.*|^BB.*|^By.*|^Bz.*")
& case ==0
browse if regexm(medcode, "^68\..*|^681.*|^686.*|^68P.*|^68Q.*|^68W.*|^68Z.*") &
case ==0
```

Table 32: Read code stems of relevance search for bowel cancer diagnosis in Stata.

The relevant child stems from this search were then browsed for inclusion. All possible relevant Read codes were extracted into Excel and sorted by Medcode. The key word search was then adjusted to include any additional terms/synonyms which may have been initially missed but identified by analysing the hierarchies. This cycle of key word search followed by hierarchical inclusion continued until all key terms had been identified along with the corresponding hierarchies.

This resulting Read code list was then subjected to review by one individual to determine which terms should be included along with the relevancy to the research question. The decision of the reviewer was marked clearly in an Excel spreadsheet. This list was then provided to a clinician for a second review before discussing the final list, providing a form of code set validation (See **Appendix 2** for 1st and 2nd reviewer decisions). Read code frequency tables provided by THIN in ancillary tables were used to determine how often a Read code was used to aid with decision making.

The final consensus Read code list had 42 codes and was then used to extract bowel cancer diagnoses from the THIN database for analysis (**Appendix 2**). This strategy ensured that all key terms were identified for extraction with clinician input.

6.0 Compiling Drug Code Lists

The drug code dictionary in THIN contains encrypted drug codes, generic drug names as well as the BNF Chapters the drug may be mapped to. This was used to produce drug code lists for use in **Chapter 5** and included; antispasmodics, anti-motility drugs and laxatives. It was investigated further as to whether these prescriptions also related to the corresponding symptom and were combined into one variable where appropriate. This was the case for abdominal pain and antispasmodic prescription since they had very similar hazard ratios.

The drug code strategy built upon previous methodology described by Davé and Petersen[3] as well as the THIN Data Guide for Researchers produced by IMS Health[1].

6.1 Drug Code List Development for Laxatives

This section describes the methods used to derive a laxative drug code list. These prescriptions can be used as a proxy for a symptom of constipation.

To formulate a key word search using the generic name listed in the drug code dictionary, the British National Formulary (www.bnf.org) was used to identify that laxative drugs were listed under Chapter 1.6. There are 4 main types of laxative: Bulk-forming laxative (1.6.1), stimulant laxatives (1.6.2), faecal softeners (1.6.3) and osmotic laxatives (1.6.4). Since the drug codes and generic drug names are mapped to BNF codes this was then used to generate the search strategy along with clinician input.

Other drug codes for laxative drugs may be listed under other BNF Chapters. In addition, over time new drugs are introduced and old drugs are removed and may be mapped to different BNF Chapters as the BNF is published every 6 months. The key word search will be able to identify these additional drug codes.

The key word search code used in Stata is shown in **Table 33**.

```
replace case=1 if
regexm(lcase,"ispaghula|fibrelief|fybogel|isogel|ispagel|regulan|methylcellulose|cele
vac|sterculia|normacol")

replace case=1 if
regexm(lcase,"bisacodyl|sodium.*picosulfate|dulcolax|pico.*liquid|pico.*perles|docusa
te.*sodium|dioctyl.*sodium.*sulphosuccinate|dioctyl|docusol|norgalax.*micro-
enema|glycerol|gylcerin|senna|sennoside|manevac|senokot|sodium.*picosulfate|dulcolax"
)

replace case=1 if regexm(lcase,"arachis.*oil|liquid.*paraffin")

replace case=1 if regexm(lcase,
"lactulose|macrogol|polyethylene.*glycol|laxido|molaxole|movicol|norgine|magnesium.*s
alt|magnesium.*hydroxide|magnesium.*sulphate|phosphates.*rectal|carbalax|sodium.*acid
.*phosphate|sodium.*dihydrogen.*phosphate|fleet|casen.*fleet|phosphates.*enema|sodium
.*citrate|microlette.*micro.*enema|micralax|relaxit")
```

Table 33: Iterative Search Strategy performed in Stata for laxative drugs.

All drug codes in Chapter 1.6 were added to the growing code list to ensure all potentially relevant codes had been added. The Stata code to add all the drug codes in Chapter 1.6 is provided in **Table 34** below.

```
replace case=1 if (regexm(bnfcode1, "^01\.06\.00")|regexm(bnfcode2,
"^01\.06\.00")|regexm(bnfcode3, "^01\.06\.00"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.01")|regexm(bnfcode2,
"^01\.06\.01")|regexm(bnfcode3, "^01\.06\.01"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.02")|regexm(bnfcode2,
"^01\.06\.02")|regexm(bnfcode3, "^01\.06\.02"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.03")|regexm(bnfcode2,
"^01\.06\.03")|regexm(bnfcode3, "^01\.06\.03"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.04")|regexm(bnfcode2,
"^01\.06\.04")|regexm(bnfcode3, "^01\.06\.04"))& case ==0
```

Table 34: Stata code to add the drugs from Chapter 1.6 to the growing drug code list for laxatives.

Any mention of dantron was removed as advised by the clinician (this drug is for terminally ill patients) along with certain types of drug formulation which are detailed in the drug code dictionary. For example, dressings, paediatric medications and creams were all excluded. This reduced the final list to review (see **Table 35**).

```
//Remove any mention of dantron as this is for terminally ill patients
replace case=0 if regexm(lcase, "dantron|co.*danthramer|co.*danthrusate")

//Scan through the types of drugs genericname
//Remove drugs which are creams, dressings, paediatric medications etc
replace case=0 if regexm(lcase,
"dressing|poultice|eye|paediatric|injection|ear.*drop|cream|biscuit|syringe|ointment|
bath.*additive|bath.*oil|emollient|soap|shampoo")

//Can we remove any by formulation?
tab formulation
replace case=0 if regexm(formulation, "dressings|drops|paediatric|infant
suppositories")
```

Table 35: Stata code to remove formulations or drugs which are not of interest for the drugcode list review for laxatives.

It was then investigated whether the drug codes identified were mapped to another Chapter by tabulating the combinations of bnfcodes included in the drug code dictionary (**Table 36**). There were 71 combinations, with Chapter 1.6 being the most frequently used. Several of the drug codes identified from the search mapped to Chapter 13 (Skin) and Chapter 7 (genito-urinary system). These were removed after the key word search but before re-adding all the drug codes mapped to Chapter 1.6 to ensure no codes in BNF Chapters 1.6 were missed (Chapter 1.6 could be in bnfcode1, bnfcode2 or bnfcode 3 column). Stata code shown below in **Table 37**.

bnfcode1	bnfcode2	bnfcode3	Combination Frequency
01.06.02.00	01.01.01.00	01.06.03.00	1
12.03.04.00	12.03.01.00	00.00.00.00	1
09.05.01.03	00.00.00.00	00.00.00.00	1
09.04.04.02	00.00.00.00	00.00.00.00	1
13.10.05.00	00.00.00.00	00.00.00.00	1
13.07.00.00	00.00.00.00	00.00.00.00	1
12.03.03.00	00.00.00.00	00.00.00.00	1
01.03.01.00	00.00.00.00	00.00.00.00	1
01.01.01.00	01.06.04.00	00.00.00.00	1
13.11.05.00	00.00.00.00	00.00.00.00	1
12.03.05.00	00.00.00.00	00.00.00.00	1
04.05.01.00	00.00.00.00	00.00.00.00	1
13.02.01.00	13.02.01.01	00.00.00.00	1
23.01.00.00	00.00.00.00	00.00.00.00	1
09.04.01.00	00.00.00.00	00.00.00.00	1
01.06.04.00	13.01.01.00	00.00.00.00	1

01.01.01.00	01.01.00.00	00.00.00.00	1
13.02.01.01	00.00.00.00	00.00.00.00	1
13.15.00.00	00.00.00.00	00.00.00.00	1
07.06.00.00	00.00.00.00	00.00.00.00	1
09.06.04.00	00.00.00.00	00.00.00.00	1
24.00.00.00	00.00.00.00	00.00.00.00	1
01.06.00.00	01.06.04.00	00.00.00.00	2
01.04.01.00	00.00.00.00	00.00.00.00	2
05.05.01.00	01.06.02.00	05.05.01.00	2
02.12.00.00	00.00.00.00	00.00.00.00	2
01.06.05.00	01.06.04.00	00.00.00.00	2
13.05.02.00	00.00.00.00	00.00.00.00	2
01.02.01.00	00.00.00.00	00.00.00.00	2
01.01.02.02	00.00.00.00	00.00.00.00	2
01.07.01.00	00.00.00.00	00.00.00.00	2
05.01.12.00	00.00.00.00	00.00.00.00	2
01.02.00.00	00.00.00.00	00.00.00.00	2
01.00.00.00	01.01.00.00	00.00.00.00	2
01.06.02.00	01.06.01.00	01.06.02.00	2
01.06.03.00	01.06.03.00	00.00.00.00	2
01.01.02.01	00.00.00.00	00.00.00.00	2
13.04.00.00	00.00.00.00	00.00.00.00	3
01.06.04.00	01.01.01.00	00.00.00.00	3
01.06.00.00	01.06.02.00	00.00.00.00	3
04.05.01.00	01.06.01.00	00.00.00.00	3
01.06.02.00	01.06.02.00	01.06.01.00	3
01.06.04.00	01.06.05.00	00.00.00.00	3
01.01.02.00	00.00.00.00	00.00.00.00	3
01.01.01.03	00.00.00.00	00.00.00.00	3
07.04.04.00	00.00.00.00	00.00.00.00	3
01.06.00.00	00.00.00.00	00.00.00.00	3
13.01.00.00	00.00.00.00	00.00.00.00	3
01.06.05.00	01.06.05.00	01.06.02.00	4
01.02.00.00	01.06.01.00	01.02.00.00	4
09.05.02.01	00.00.00.00	00.00.00.00	4
01.01.01.01	00.00.00.00	00.00.00.00	5
01.02.00.00	01.02.00.00	01.06.01.00	5
01.06.04.00	01.06.03.00	01.01.01.00	5
13.09.00.00	00.00.00.00	00.00.00.00	6
13.02.01.02	00.00.00.00	00.00.00.00	6
03.09.01.00	00.00.00.00	00.00.00.00	6
01.01.01.00	01.01.01.00	00.00.00.00	7
03.09.02.00	00.00.00.00	00.00.00.00	7
01.06.05.00	00.00.00.00	00.00.00.00	8

01.01.01.00	00.00.00.00	00.00.00.00	9
13.02.01.00	00.00.00.00	00.00.00.00	10
01.06.01.00	01.06.01.00	00.00.00.00	10
01.06.03.00	00.00.00.00	00.00.00.00	13
07.04.03.00	00.00.00.00	00.00.00.00	19
01.06.04.00	01.06.04.00	00.00.00.00	22
99.00.00.00	00.00.00.00	00.00.00.00	23
00.00.00.00	00.00.00.00	00.00.00.00	41
01.06.01.00	00.00.00.00	00.00.00.00	83
01.06.04.00	00.00.00.00	00.00.00.00	119
01.06.02.00	00.00.00.00	00.00.00.00	127

Table 36: Combination frequencies of bnfcodes included in the drug code dictionary for the drug code list for review for laxatives.

```
replace case=0 if regexm(bnfcode1, "^13\..*|^07\..*")
```

Table 37: Stata code to remove drug codes mapped to Chapters which are not of relevance to the final drugcode list for review for laxatives

The Anatomical Therapeutic Chemical (ATC) classification system can be used as a final check to assess drug ingredients but an ATC code is not present for all records. These are alphanumeric codes developed by the World Health Organisation to classify drugs. This would involve searching for laxative drugs under ATC codes beginning with A06A* (drugs for constipation). Drug code frequency tables provided by THIN in ancillary tables were used to determine how often a certain drug code was used to aid with decision making.

This final list was subjected to a first review and then for validity by a second reviewer with clinical expertise in this area. The consensus drug code list had 450 codes (**Appendix 3**).

7.0 DISCUSSION

7.1 Statement of Principle Findings

This chapter has presented the methodology used to define acceptable periods of BCSP notifications for practices receiving electronic results (AEB date). The initial review of the AEB date defined for each practice identified 353 practices for inclusion and 102 for exclusion. The second reviewer gave 363 for inclusion and 92 practices for exclusion which is a 97.8% (445/455) agreement. Eighty percent of practices were investigated for an AEB date. In 2017, NHS Digital suggested that 88% of GP practices receive electronic BCSP notifications (NHS Digital, Personal Communication). The difference of 8% may be explained since this study used THIN May 2016 Version so it might be that practices have updated their notification systems since this time. Reasons for exclusion of practices was if they had a zero rate of notifications received or extremely low numbers of irregular peaks or if they had too short a duration before they stopped contributing to THIN which may make up the missing 8% (about 36 practices). In addition, some practices were excluded where it was not clear when the electronic notifications commenced. The different patterns in electronic notifications seen by the practices can be influenced by practice behaviour, patient behaviour, the sending of lab results and the change/update in IT Systems such as Vision software.

This chapter also detailed the methods used to extract AHD variables using the examples of BCSP notifications and a lab test result of haemoglobin concentration. The AHD file is more complex in structure compared to the Medcode file and therefore requires more investigation before data extraction. Generating an ahdcode list (like with Read codes and Drug code extraction) would not be sufficient to extract the level of detail required for a study.

For BCSP outcomes this approach involved using Read codes for BCSP notifications with a definitive result from the AHD file. Other generic BCSP codes however needed to be combined with other ahdcodes for a definitive outcome/result. For haemoglobin concentration, the method sequentially reviewed the numeric distribution of Hb concentration values for combinations of Read codes and unit value labels observed in a THIN 1% sample. These were compared with a reference distribution for Hb concentration to assess if they matched that distribution, required conversion before use, or were

unlikely to contain Hb values and therefore excluded. Plausible Hb minimum and maximum values from an external source were then applied to the data as a final step in the method.

The methods used to compile Read code lists were presented using the example of 'bowel cancer diagnoses'. Although previous studies or clinical code repositories may have the associated Read codes presented alongside the paper, Read code lists need to be tailored to each research question and study. The outcome of interest in **Chapter 5** was colorectal cancer or polyp diagnosis and so the methods to derive code lists for bowel cancer diagnosis was described. Read code list generation involved an iterative key word search strategy (informed by previous research and discussion with a clinician) and then identifying other parent and child stems for Read codes by exploiting their hierarchical nature. After a few cycles of running through this method, the resulting Read code list was subject to two individual reviewers (the second being a clinician for validity). The final consensus list was made up of 42 codes for bowel cancer diagnosis.

The method of drug code list development is presented for laxative prescriptions. Drug code list generation involved a similar method to the above. The British National Formulary was used to identify Chapters the drugs were mapped to as well as formulate a key word search under generic drug name. All drugs mapped to the appropriate Chapter in the BNF were included in the final list after removing any drugs which did not fit the definition used for the study reported in **Chapter 5** (e.g. formulations or other indications). The list was subjected to two reviewers. The final drug code list for laxatives resulted in a list of 450 codes.

7.2 Strengths and Weaknesses

This study is the first to present the methodology used to define an AEB date for research studies using BCSP data from a Primary Care Database. **Chapter 5** used the AEB date to help define a bowel cancer screening programme cohort for analysis as well as to provide validity to the data.

For defining the AEB date, a common x and y axis was used to aid comparison of the graphs and to identify the change point. Validity of the chosen AEB date for each practice was enhanced by having a second reviewer and consensus meeting where a 3rd reviewer contributed where necessary. The approach of using pairs of reviewers to assign dates was also used by the study which developed the AMR date.[14]

A limitation of the AEB date development was that the AEB date identified for each practice could be considered subjective which can introduce bias. A set of rules for inclusion and exclusion of practices as well as identifying the date was derived for use by both the 1st and 2nd reviewer to improve consistency. There was strong agreement in the consensus meeting for these dates. In addition to this, a reference line and LOWESS were plotted alongside the actual notification rate to assist with decision making. Another approach to visual interpretation is using computer derived change point analysis to auto-categorise the practice plots. However, due to the variety of patterns seen in the plot outputs this was not feasible.

A key strength of this chapter is the level of detail which is presented to aid the reproducibility of the research. Reporting guidelines such as the RECORD statement[23] recommend that full clinical code lists and algorithms are presented alongside the research (RECORD 7.1). For instance, clinical codes could also be published alongside the journal article or placed in a data repository such as ClinicalCodes[22]. More recently it has been suggested that this perhaps does not go far enough and that researchers should also provide some form of 'meta-data' to go alongside the clinical code lists so other researchers can critique, amend or apply similar methods in future research.[2] For instance, it has been suggested that the initial set of synonyms used to search for the clinical codes could be included.[2] Related to this, is using the statistical software Stata to develop these methods since do-files enhance replication and reproducibility. Stata code is presented within the chapter to aid the method to be adapted or used in further research.

Code set validity is strengthened by having an expert clinician as a second reviewer and to assist with compiling key words for the search strategy. This approach has been recommended by a review of clinical code set engineering to reduce Type I errors (where a code is wrongly included).[2] It has also been suggested that this approach could be extended further by using an expert panel to decide on the final clinical code lists.[2 24] Furthermore, by using a broad set of synonyms, building on previous code lists, exploiting the hierarchical nature of codes and following an iterative process (by including additional synonyms based on codes discovered by searching code hierarchy) this helps to reduce type II errors (where a code is incorrectly excluded).[2]

By defining both prescriptions and symptoms, the related prescription could be used as an indicator for disease in **Chapter 5**. For instance, antispasmodic prescription was found to have a similar hazard ratio to the symptom of abdominal pain and this was combined into

one variable. Other studies have also investigated using laxatives as a proxy to a constipation symptom and anti-motility drugs as a proxy to a symptom of diarrhoea.[21 25] This approach increased the sensitivity of code lists and identification of a particular clinical feature.

When devising a search strategy for Read code and Drug code list development, the 'regexm' (regular expressions) function was used in Stata.[26] This function is more flexible than the 'strpos' function,[27] for building a keyword search strategy and allows different combinations of words to be searched as well as truncated forms. The 'strpos' function was used as an example in the study by Davé and Petersen[3] which reports methods for developing medical and drug code lists. The more flexible 'regular expressions' approach was used in the current study to build an extensive search strategy.

Another key strength when developing a strategy to extract lab data from the AHD file is the use of a reference distribution obtained from a local hospital laboratory for GP requested Hb tests. To ensure the distribution of lab test results are consistent with other datasets and to remove potential outliers from analysis, this external local hospital dataset was used to set minimum and maximum Hb values to the data extracted from THIN. This improves the validity of the data further and the associations found in **Chapter 5**.

Olier *et al.*[24] present the methodology used to develop a list of clinical codes and provide accompanying Stata and R commands. This method is applicable to different coding systems which has the added advantage of being used for SNOMED CT which will be implemented in all practices by the end of 2018.[4] Although the methodology used in the current study is specific to Read code Version 2 which is used in the THIN database, the principles are equally valid for new coding systems. At the time of planning and carrying out the study in **Chapter 5**, THIN used Read code Version 2 for documenting clinical signs and symptoms. In addition, database providers will provide interim tables for codes to enable this transition and the use of Read code lists.

7.4 Practical implications

The methods used for defining the AEB date and the resultant dates extracted for each practice can be used in future studies using BCSP data. The AEB date can be used to provide quality assurance to the data as well as to help define a screening cohort for analysis from data which is principally used for primary care based studies. This is the approach taken for the study reported in **Chapter 5**.

Although these methods were developed in THIN a similar approach can be applied in other electronic health care databases such as the CPRD and QRESEARCH as a data quality assurance filter for BCSP studies. Further to this, electronic screening programme notifications for other national programmes such as breast cancer and cervical screening can be investigated to define screening cohorts for analysis and for data quality assurance. Different regions in the UK have separate IT systems to record screening activities for their regional screening programmes. The methods can therefore also be applied to practices in Scotland, Ireland, Wales and the Isle of Man.

Although there have been several studies which report the methodology of Read code and Drug code list development,[3 8 24 28] as far as can be identified no studies have reported their methods for AHD variable extraction. This is a much more complex data structure and it is perhaps even more paramount to detail the methods used to derive valid data, particularly when involving quantitative measures such as lab test results. The methods reported in this chapter can be used when defining new variables for use in the THIN database.

The Read code and Drug code lists along with the methods for their construction can be used or amended for use to fit a particular research question by other researchers in future studies. Any publications from this Chapter and **Chapter 5** will publish the clinical code lists (as well as other forms of meta-data) alongside the journal article as recommended by reporting guidelines and reviews.[2 23]

Further investigations into the temporal changes in code recording for clinical indicators could be investigated. For instance, with the introduction of QOF incentives or changes in NICE guidelines, there would be a corresponding effect on what is recorded for a particular condition. The usage of one particular Read code could reduce over time with the increase in the use of another. The odds ratios (or hazard ratios) could also be analysed over a certain timespan to determine whether there is a change in recording or change in the prevalence/incidence rates of a disease.

7.5 Future research

The methods used to derive the AEB date in this chapter could be transferred for use in identifying acceptable lab test reporting dates (from pathology labs) since the BCSP use the same system used by NHS pathology labs to send results through to GP surgeries. Pathology labs would have switched over from paper based records to electronic notifications and so a similar approach to the one reported here could be used as a layer of

quality assurance to the data extracted. This is because it is more likely that data before electronic notifications would have been less complete, as well as biased towards reporting abnormal results.

As clinical reporting switches over to SNOMED CT in the NHS, the methods reported in this study can be adapted for use with this clinical terminology system. A standardised approach such as the three-stage process reported by Watson *et al.*[8] which does not rely on the hierarchical structure of Read codes, would be more appropriate in this circumstance for SNOMED CT code list development. For instance, sources of clinical information which could be used to help define the clinical feature of interest include, ICD10, BMJ Best Practice Guidelines, International Classification of Primary Care (ICPC) and NICE Clinical Knowledge Summaries.[8] The three-stage process also includes a modified Delphi approach with primary care practitioners to reach a consensus of the most relevant codes.[2]

A quantitative Faecal Immunochemical test code exists for SNOMED CT,[29] it is unknown at this stage whether these codes will be adopted by the NHS BCSP or whether the more generic FOBT codes will be retained. If the quantitative FIT amount is sent to GPs this level of information can be used in a variety of ways including for research, to indicate the level of risk to an individual which may help uptake[30] and to determine longitudinal changes to FIT.

8.0 CONCLUSIONS

The AEB date can be used in future studies as a layer of data quality assurance to ensure results used for analysis are ones which have been electronically received to the additional health data records. The AEB date can also help to derive an average risk BCSP cohort for analysis from electronic primary care records by limiting the population to the age range for screening 60-74 and excluding those with high risk familial syndromes. With the increasing popularity of electronic health record research, it is essential that the methods used to compile clinical code lists as well as extracting additional health data are rigorous and reproducible in order to produce valid results. The methods described for additional health data variables such as haemoglobin concentration can be used for future studies or amended for other types of lab test result. By using an external dataset to define acceptable minimum and maximum values this also adds more validity to the data when applied in risk prediction models. There is a push to ensure Read code / Drug code lists are

transparent and available for future research studies which ties into the trend for reproducible research (for instance through repositories such as ClinicalCodes). The definition of diagnoses and symptoms using these Read codes need to be specific to the research question and the researcher must tailor lists for their own use. Changes in the clinical terminology system used in the NHS in the form of SNOMED CT will require an adaptation of these methods but using similar principles to those reported here. This chapter has described the methods developed to extract data for the study reported in **Chapter 5** but also which can be applied in other future studies to improve data validity and quality assurance.

The final chapter is the thesis discussion (**Chapter 7**) presenting a summary of the findings from each chapter and the corresponding future perspectives in this area of research.

9.0 REFERENCES

1. IMS Health. THIN Data Guide for Researchers 3.0. 2015.
2. Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. Journal of Biomedical Informatics. 2017;70:1-13.
3. Davé S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiology and Drug Safety. 2009;18(8):704-7.
4. NHS Digital. SNOMED CT implementation in primary care 2018 [Available from: https://digital.nhs.uk/SNOMED-CT-implementation-in-primary-care.
5. NHS England. National Pathology Programme. Digital First: Clinical Transformation through Pathology Innovation 2014 [Available from: https://www.england.nhs.uk/wp-content/uploads/2014/02/pathol-dig-first.pdf.
6. NHS Connecting for Health. Bowel Cancer Screening Programme - Electronic Results Communication 2010 [Available from: http://www.inps.co.uk/my-vision/user-guides-downloads/user-guides/nhs-bowel-cancer-screening-frequently-asked-questions.
7. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015;12(10):e1001885.
8. Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open. 2017;7(11).
9. NHS Digital. Quality and Outcomes Framework (QOF) online database 2018 [Available from: https://digital.nhs.uk/QOF-online-database.
10. Taggar JS, Coleman T, Lewis S, Szatkowski L. The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: cross-sectional analyses from The Health Improvement Network (THIN) database. BMC Public Health. 2012;12(1):329.
11. Kontopantelis E, Reeves D, Valderas JM, Campbell S, Doran T. Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study. BMJ Quality & Safety. 2013;22(1):53-64.
12. Kontopantelis E, Olier I, Planner C, Reeves D, Ashcroft DM, Gask L, et al. Primary care consultation rates among people with and without severe mental illness: a UK cohort study using the Clinical Practice Research Datalink. BMJ Open. 2015;5(12).
13. NHS Digital. NHS e-Referral Service 2018 [Available from: https://digital.nhs.uk/e-Referral-Service.
14. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. Pharmacoepidemiol Drug Saf. 2009;18(1):76-83.
15. Horsfall L, Walters K, Petersen I. Identifying periods of acceptable computer usage in primary care research databases. Pharmacoepidemiol Drug Saf. 2013;22(1):64-9.
16. Logan RF, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. Gut. 2012;61(10):1439-46.
17. INPS (In Practice Systems Ltd) Vision. Bowel Cancer Screening England 2013 [Available from: http://www.inps.co.uk/my-vision/user-guides-downloads/user-guides/regional-user-guides/england/bowel-cancer-screening.
18. Berg J. The UK Pathology Harmony initiative; The foundation of a global model. Clinica Chimica Acta. 2014;432:22-6.

19. Provan D, Baglin T, Dokal I, de Vos J. Oxford Handbook of Clinical Haematology. 4th ed. New York, United States: Oxford University Press; 2015.
20. Provan DB, Trevor, Dokal I, de Vos J. Oxford Handbook of Clinical Haematology. 4th ed. New York, United States: Oxford University Press; 2015.
21. Marshall T, Lancashire R, Sharp D, Peters TJ, Cheng KK, Hamilton W. The diagnostic performance of scoring systems to identify symptomatic colorectal cancer compared to current referral guidance. Gut. 2011;60(9):1242-8.
22. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, et al. ClinicalCodes: An Online Clinical Codes Repository to Improve the Validity and Reproducibility of Research Using Electronic Medical Records. PLoS ONE. 2014;9(6):e99825.
23. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. 2015;12(10):e1001885.
24. Olier I, Springate DA, Ashcroft DM, Doran T, Reeves D, Planner C, et al. Modelling Conditions and Health Care Processes in Electronic Health Records: An Application to Severe Mental Illness with the Clinical Practice Research Datalink. PLoS One. 2016;11(2):e0146715.
25. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng K, Marshall T. The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. BMC medicine. 2009;7:17.
26. Stata.com. Stata 15 help for regexm() 2018 [Available from: https://www.stata.com/help.cgi?regexm().
27. Stata.com. strpos() - Find substring in string 2018 [Available from: https://www.stata.com/manuals13/m-5strpos.pdf.
28. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toschke AM, for the e CRTRT. Selection of Medical Diagnostic Codes for Analysis of Electronic Patient Records. Application to Stroke in a Primary Care Database. PLOS ONE. 2009;4(9):e7168.
29. National Pathology Exchange. SNOMED CT Browser 2018 [Available from: https://snomedbrowser.com/Codes/Details/1032221000000105.
30. Edwards AG, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. The Cochrane database of systematic reviews. 2013(2):Cd001865.

10.0 APPENDICES

Appendix 1 – Acceptable Electronic BCSP (AEB) date derived for each practice using THIN (Version May 2016)

Practice ID (pseudonymised)	Screening Start Date	Practice Inclusion
0036	01oct2011	Y
0242	01apr2013	Y
0341		N
0267		N
0333	01apr2013	Y
0258	01mar2011	Y
0255	01jul2010	Y/S
0235		N
0298	01jan2010	Y
0083	01jul2009	Y
0264	01sep2009	Y
0351	01jul2012	Y
0275	01jul2013	Y
0285	01sep2012	Y
0383		N
0214	01jun2011	Y
0361	01mar2013	Y
0400	01feb2013	Y
0381		N
0382	01nov2010	Y
0376	01sep2012	Y
0136	01apr2013	Y
0014		N
0343	01apr2011	Y
0311		N
0002	01jun2011	Y
0416	01dec2010	Y
0031		N
0165	01oct2013	Y
0399	01sep2012	Y
0070	01sep2011	Y/S
0110	01jun2013	Y
0247	01jan2011	Y
0419	01aug2013	Y
0377		N
0004	01jul2011	Y
0388	01nov2010	Y
0115	01jan2012	Y
0132	01dec2013	Y

0335	01sep2012	Y
0260	01feb2011	Y
0248	01nov2010	Y
0003	01oct2012	Y
0370		N
0398	01jul2013	Y/S
0037	01mar2014	Y
0405	01nov2011	Y/S
0237		N
0200	01mar2011	Y
0100	01jul2013	Y/S
0257		N
0010	01oct2010	Y
0211		N
0071	01nov2010	Y
0429	01mar2011	Y
0321		N
0287		N
0159		N
0130		N
0174		N
0407	01may2014	Y
0041	01oct2010	Y
0180		N
0276	01jan2011	Y
0296	01oct2012	Y
0114	01may2009	Y
0229	01sep2012	Y
0409	01jul2009	Y
0436	01apr2014	Y
0234		N
0355	01feb2012	Y
0326	01jul2014	Y
0413	01may2014	Y
0086	01may2013	Y
0271	01jun2013	Y
0395	01oct2010	Y
0019		N
0313	01apr2013	Y
0081	01jul2013	Y
0446	01jul2012	Y
0072	01feb2011	Y/S
0125	01jul2011	Y
0282	01apr2011	Y

0449	01apr2011	Y
0091		N
0306		N
0431	01dec2010	Y
0327	01jul2012	Y
0365		N
0001	01feb2011	Y
0228		N
0290	01nov2011	Y
0314	01jun2009	Y
0256		N
0278	01aug2011	Y
0101	01oct2013	Y
0181	01aug2013	Y
0088	01mar2013	Y
0025	01feb2011	Y
0149	01sep2013	Y
0273	01nov2010	Y
0195		N
0198	01jul2013	Y
0140	01oct2010	Y
0384	01apr2011	Y
0309	01sep2012	Y
0065	01mar2011	Y
0084	01jul2009	Y
0455		N
0337	01feb2011	Y
0433	01jul2013	Y
0166		N
0364	01jul2011	Y
0283	01jul2012	Y
0344	01nov2011	Y
0162	01oct2014	Y
0240	01apr2014	Y
0020		N
0015	01jun2012	Y
0270	01jun2011	Y
0151	01jul2010	Y
0170		N
0318	01jan2013	Y
0373		N
0203	01nov2010	Y
0356		N
0139	01apr2013	Y

0060	01apr2011	Y
0277	01jul2011	Y
0378	01may2010	Y
0334	01may2011	Y
0008	01jan2011	Y
0263	01sep2012	Y
0108		N
0207	01jan2013	Y
0048		N
0026	01apr2011	Y
0122	01sep2012	Y
0360	01jul2013	Y
0011	01sep2014	Y
0426	01apr2014	Y
0225	01jan2012	Y
0386	01jun2013	Y
0389	01apr2013	Y
0216	01jan2013	Y
0182	01sep2014	Y
0423	01apr2013	Y
0164	01jun2013	Y
0171	01jan2011	Y
0117	01may2013	Y
0080	01may2011	Y
0411	01feb2013	Y
0250	01may2013	Y/S
0281		N
0294	01mar2013	Y
0199	01aug2012	Y
0241	01nov2013	Y/S
0385		N
0051	01aug2012	Y
0322	01sep2014	Y
0412		N
0190	01may2011	Y
0030	01apr2013	Y
0319		N
0112	01feb2011	Y
0286	01mar2011	Y
0116	01jun2012	Y
0292		N
0268		N
0372	01aug2013	Y/S
0201	01aug2012	Y

0067	01apr2011	Y
0178	01dec2010	Y
0017	01dec2010	Y
0252		N
0254	01nov2012	Y
0236	01nov2010	Y
0043	01nov2010	Y
0092		N
0293	01may2013	Y
0274	01nov2010	Y
0121	01feb2013	Y/S
0284	01nov2014	Y
0435		N
0055	01mar2013	Y
0176	01mar2011	Y
0075		N
0315	01aug2013	Y
0147	01jul2015	Y
0193		N
0039	01jul2012	Y
0059	01aug2013	Y
0049		N
0047	01nov2013	Y
0336	01jul2011	Y
0097	01feb2013	Y
0050	01mar2011	Y/S
0454	01feb2011	Y
0111	01jan2013	Y
0445	01mar2014	Y
0024	01may2013	Y
0208	01sep2013	Y
0391	01dec2011	Y
0167	01apr2011	Y
0146	01jan2012	Y
0066		N
0157	01jul2012	Y
0332	01nov2014	Y
0219	01oct2012	Y
0408	01jul2014	Y/S
0076	01feb2011	Y
0403	01oct2012	Y
0406	01dec2014	Y
0302	01feb2011	Y
0052	01mar2012	Y

0161	01mar2013	Y
0440		N
0244	01may2011	Y
0224	01feb2013	Y
0380		N
0123	01oct2012	Y
0099		N
0415		N
0129	01nov2010	Y
0295	01feb2011	Y
0153	01mar2011	Y
0401	01jul2012	Y
0077	01apr2013	Y
0231	01sep2012	Y/S
0194	01dec2011	Y
0329	01jan2012	Y
0350	01sep2014	Y
0452	01apr2014	Y
0202	01aug2009	Y
0213	01may2009	Y
0450	01jan2013	Y
0299	01jan2014	Y
0196		N
0434	01jul2012	Y
0184		N
0245	01dec2012	Y
0102		N
0451	01apr2012	Y
0186	01jun2011	Y
0369	01jun2012	Y
0441		N
0104	01nov2013	Y
0197		N
0118	01sep2013	Y
0044	01jul2011	Y
0185	01jul2011	Y/S
0414	01apr2014	Y
0392	01feb2011	Y
0068		N
0107	01mar2013	Y
0188	01may2014	Y
0144	01sep2012	Y
0430	01jul2012	Y
0262	01jul2011	Y

0128		N
0007	01aug2013	Y
0058	01jul2012	Y
0300	01feb2012	Y
0425	01jun2011	Y/S
0253	01dec2009	Y
0141		N
0308	01may2011	Y
0347	01aug2013	Y
0339	01nov2010	Y
0363	01mar2011	Y
0218		N
0046	01nov2009	Y
0305	01jul2012	Y
0142	01nov2010	Y
0418		N
0062	01jun2011	Y
0127	01may2011	Y
0362		N
0374		N
0307	01may2012	Y
0045		N
0089	01oct2012	Y
0346	01apr2011	Y
0143		N
0227	01oct2013	Y/S
0150	01mar2010	Y
0172	01oct2013	Y
0368	01jul2011	Y
0133	01oct2013	Y
0222	01mar2014	Y
0352	01feb2011	Y
0016		N
0421	01oct2014	Y
0331	01may2012	Y
0422	01dec2010	Y
0085	01apr2013	Y
0265		N
0297		N
0272		N
0338	01jul2011	Y
0138	01sep2011	Y
0183	01jul2012	Y
0191		N

0353	01mar2013	Y
0221		N
0090	01jun2012	Y
0444	01aug2013	Y
0158	01nov2013	Y/S
0145		N
0056	01jan2012	Y
0371		N
0074	01nov2011	Y
0324	01may2011	Y
0291	01jun2014	Y
0134	01feb2011	Y
0217	01aug2011	Y
0009	01sep2011	Y
0021	01may2014	Y
0087	01may2013	Y
0173	01mar2011	Y
0266		N
0410	01dec2010	Y
0233	01apr2011	Y
0033	01feb2015	Y
0328	01feb2011	Y
0029		N
0192	01oct2012	Y
0312		N
0154	01sep2012	Y
0325	01apr2012	Y
0005		N
0053		N
0246	01sep2014	Y
0342	01oct2013	Y/S
0120	01aug2013	Y
0249	01jan2013	Y/S
0063	01may2011	Y
0288	01apr2011	Y
0387	01nov2011	Y
0187	01apr2012	Y
0230	01dec2012	Y
0358	01oct2011	Y
0105	01jul2011	Y
0243	01mar2011	Y
0301	01oct2012	Y
0119	01apr2014	Y/S
0303	01nov2015	Y/S

0289	01apr2011	Y
0006	01mar2013	Y
0027	01may2011	Y
0417		N
0054	01jul2012	Y
0354	01may2014	Y
0379	01aug2013	Y
0320	01feb2011	Y
0124	01feb2014	Y
0367		N
0135	01mar2011	Y
0098	01feb2011	Y
0345	01sep2014	Y
0018	01dec2010	Y
0113	01oct2010	Y
0156	01aug2010	Y
0420	01feb2013	Y
0238		N
0439	01oct2013	Y/S
0261	01may2013	Y
0348	01feb2014	Y
0442	01may2009	Y
0169	01nov2010	Y
0126	01oct2012	Y
0095	01may2013	Y
0082	01feb2014	Y
0152		N
0013	01apr2014	Y
0340	01apr2014	Y/S
0109	01nov2014	Y
0232		N
0357	01apr2011	Y
0148	01nov2014	Y
0177	01nov2012	Y
0453	01sep2010	Y/S
0096	01nov2013	Y
0204	01apr2011	Y
0393	01nov2010	Y
0073	01nov2013	Y
0280	01sep2013	Y
0206	01aug2012	Y
0093		N
0057	01may2013	Y/S
0226	01mar2011	Y

0437	01feb2011	Y
0279	01aug2013	Y
0223	01may2014	Y
0304	01jan2011	Y
0432	01aug2013	Y
0310	01dec2012	Y
0212	01jan2014	Y
0131	01nov2014	Y
0061	01feb2012	Y
0448	01mar2011	Y
0023	01dec2013	Y
0042	01apr2013	Y
0404	01nov2010	Y
0447	01nov2010	Y
0366	01nov2013	Y
0012	01feb2011	Y
0438	01feb2014	Y
0038	01aug2012	Y
0269	01sep2014	Y
0078	01jul2011	Y/S
0064	01apr2014	Y
0163		N
0106	01nov2013	Y
0209	01oct2009	Y
0215	01mar2011	Y
0220		N
0079	01apr2011	Y
0155	01jan2014	Y
0040	01mar2011	Y
0330	01nov2014	Y
0069	01sep2011	Y
0316	01aug2012	Y
0022	01aug2013	Y
0034		N
0359		N
0394		N
0402	01aug2012	Y/S
0175	01mar2011	Y
0032	01oct2014	Y
0251	01jul2014	Y
0424	01jun2012	Y/S
0094	01jan2010	Y
0427	01jun2012	Y
0428	01jan2012	Y/S

0189	01feb2011	Y
0028	01nov2013	Y
0317	01sep2013	Y
0396	01jul2012	Y
0259	01mar2011	Y
0205		N
0103	01nov2010	Y
0397	01apr2011	Y
0168	01sep2013	Y
0323	01nov2013	Y
0137	01mar2011	Y
0239		N
0160	01dec2010	Y
0443	01may2014	Y
0210	01jul2009	Y
0035	01jan2015	Y
0375	01jul2013	Y
0349	01nov2010	Y
0179	01mar2011	Y
0390	01nov2010	Y

Table A.1.1: Acceptable Electronic BCSP (AEB) date derived for each practice using THIN (Version May 2016). Original practice ID (pracid on THIN database) and the practice region is available from the authors on request.

Appendix 2 – Read Code Lists for Bowel Cancer Diagnosis

Medcode (Bowel cancer diagnosis)	description
14CC.00	H/O Lower GIT Neoplasm
68W2400	Bowel scope (flexible sigmoidoscopy) screen: cancer detected
9Ow1.00	Bowel cancer detected by national screening programme
B13..00	Malignant neoplasm of colon
B130.00	Malignant neoplasm of hepatic flexure of colon
B131.00	Malignant neoplasm of transverse colon
B132.00	Malignant neoplasm of descending colon
B133.00	Malignant neoplasm of sigmoid colon
B134.00	Malignant neoplasm of caecum
B134.11	Carcinoma of caecum
B135.00	Malignant neoplasm of appendix
B136.00	Malignant neoplasm of ascending colon
B137.00	Malignant neoplasm of splenic flexure of colon
B138.00	Malignant neoplasm, overlapping lesion of colon

B13y.00	Malignant neoplasm of other specified sites of colon
B13z.00	Malignant neoplasm of colon NOS
B13z.11	Colonic cancer
B14..00	Malignant neoplasm of rectum, rectosigmoid junction and anus
B140.00	Malignant neoplasm of rectosigmoid junction
B141.00	Malignant neoplasm of rectum
B141.11	Carcinoma of rectum
B141.12	Rectal carcinoma
B142000	Malignant neoplasm of cloacogenic zone
B14y.00	Malig neop other site rectum, rectosigmoid junction and anus
B14z.00	Malignant neoplasm rectum,rectosigmoid junction and anus NOS
B180200	Malignant neoplasm of retrocaecal tissue
B18y000	Malignant neoplasm of mesocolon
B18y100	Malignant neoplasm of mesocaecum
B18y200	Malignant neoplasm of mesorectum
B1z0.11	Cancer of bowel
BB52000	[M]Adenocarcinoma in tubulovillous adenoma
BB5L100	[M]Adenocarcinoma in adenomatous polyp
BB5L300	[M]Adenocarcinoma in multiple adenomatous polyps
BB5M.00	[M]Tubular adenomas and adenocarcinomas
BB5M100	[M]Tubular adenocarcinoma
BB5Mz00	[M]Tubular adenoma or adenocarcinoma NOS
BB5R800	[M]Adenocarcinoid tumour
BB5U.00	[M]Villous adenomas and adenocarcinomas
BB5U100	[M]Adenocarcinoma in villous adenoma
BB5U200	[M]Villous adenocarcinoma
BB5Uz00	[M]Villous adenoma or adenocarcinoma NOS
BB83.00	[M]Pseudomyxoma peritonei

A.2.1: Final Read code list for Bowel Cancer Diagnosis.

medcode	description	JC_decision	TM_decision	Notes
122F.00	No family history of bowel cancer	0		
1241.12	FH: Bowel cancer	0		
124F.00	FH: Bowel cancer	0		
14CB.00	H/O Upper GIT Neoplasm	0		
14CC.00	H/O Lower GIT Neoplasm	1	Include	Specific enough?
1J0J.00	Suspected gynaecological cancer	0		
38GT000	QCancer colorectal cancer risk calculator	0		
6864.00	Large bowel neoplasm screen	0		
6864.11	Colon neoplasm screen	0		
6864.12	Rectal neoplasm screen	0		
6866.00	Bowel cancer screening programme: faecal occult blood result	0		
68W2.00	Bowel cancer screening programme	0		
68W2000	Bowel Cancer Screening Programme bowel scope screening test	0		
68W2400	Bowel scope (flexible sigmoidoscopy) screen: cancer detected	1	Include	
68W2500	Bowel scope (flexi-sig) screen: suspected cancer detected	0		
8CAo.00	Patient given advice about bowel cancer	0		
8Hn1.00	Fast track referral for suspected gynaecological cancer	0		
8Hn4.00	Fast track referral for suspected colorectal cancer	0		Not definitive enough
8IA3.00	Bowel cancer screening declined	0		
8OA5.00	Prov of written info about bowel cancer screening programme	0		
9Ni2.00	Did not attend bowel cancer screening programme nurse clinic	0		
9Ni3.00	Did not attend bowel cancer screening	0		
9Nic000	DNA fast track suspected gynaecological cancer clinic	0		
9Np7.00	Seen in fast track suspected colorectal cancer clinic	0		
9Ow..00	Bowel cancer screening programme administration	0		
9Ow1.00	Bowel cancer detected by national screening programme	1	Include	
9Ow2.00	No response to bowel cancer screening programme invitation	0		
9Ow3.00	Not eligible for bowel cancer screening programme	0		
9Ow4.00	Bowel cancer screening programme telephone invitation	0		
9Ow5.00	Bowel cancer screening programme invitation letter sent	0		
A4z1.00	Adenoviral meningitis	0		
B....00	Neoplasms	0		Parent code/not specific enough
B....11	Cancers	0		Parent code/not specific enough
B1...00	Malignant neoplasm of digestive organs and peritoneum	0		Not specific enough?
B1...11	Carcinoma of digestive organs and	0		Not specific enough?

	peritoneum			
B13..00	Malignant neoplasm of colon	1	Include	
B130.00	Malignant neoplasm of hepatic flexure of colon	1	Include	
B131.00	Malignant neoplasm of transverse colon	1	Include	
B132.00	Malignant neoplasm of descending colon	1	Include	
B133.00	Malignant neoplasm of sigmoid colon	1	Include	
B134.00	Malignant neoplasm of caecum	1	Include	
B134.11	Carcinoma of caecum	1	Include	
B135.00	Malignant neoplasm of appendix	1	Include	
B136.00	Malignant neoplasm of ascending colon	1	Include	
B137.00	Malignant neoplasm of splenic flexure of colon	1	Include	
B138.00	Malignant neoplasm, overlapping lesion of colon	1	Include	
B139.00	Hereditary nonpolyposis colon cancer	0	Exclude	Not including HNPCC as high risk group not average risk
B13y.00	Malignant neoplasm of other specified sites of colon	1	Include	
B13z.00	Malignant neoplasm of colon NOS	1	Include	
B13z.11	Colonic cancer	1	Include	
B14..00	Malignant neoplasm of rectum, rectosigmoid junction and anus	1	Include	
B140.00	Malignant neoplasm of rectosigmoid junction	1	Include	
B141.00	Malignant neoplasm of rectum	1	Include	
B141.11	Carcinoma of rectum	1	Include	
B141.12	Rectal carcinoma	1	Include	
B142000	Malignant neoplasm of cloacogenic zone	1	Include	Rare tumour of the anorectal region
B14y.00	Malig neop other site rectum, rectosigmoid junction and anus	1	Include	
B14z.00	Malignant neoplasm rectum,rectosigmoid junction and anus NOS	1	Include	
B180200	Malignant neoplasm of retrocaecal tissue	1	Include	
B18y000	Malignant neoplasm of mesocolon	1	Include	
B18y100	Malignant neoplasm of mesocaecum	1	Include	
B18y200	Malignant neoplasm of mesorectum	1	Include	
B1z..00	Malig neop oth/ill-defined sites digestive tract/peritoneum	0		
B1z..00	Malig neop oth/ill-defined sites digestive tract/peritoneum	0		
B1z0.00	Malignant neoplasm of intestinal tract, part unspecified	0		
B1z0.00	Malignant neoplasm of intestinal tract, part unspecified	0		
B1z0.11	Cancer of bowel	1	Include	
B1z2.00	Malignant neoplasm, overlapping lesion of digestive system	0		
B1zy.00	Malignant neoplasm other spec digestive tract and peritoneum	0		
B1zz.00	Malignant neoplasm of digestive tract and peritoneum NOS	0		

B5...00	Malignant neoplasm of other and unspecified sites	0		
B5...11	Carcinoma of other and unspecified sites	0		
B57..00	Secondary malig neop of respiratory and digestive systems	0		Secondary
B57..11	Metastases of respiratory and/or digestive systems	0		
B57..12	Secondary carcinoma of respiratory and/or digestive systems	0	Exclude	
B575.00	Secondary malignant neoplasm of large intestine and rectum	0	Exclude	Unsure whether to include secondary malignant - I suppose screening would detect these so perhaps I should?
B575000	Secondary malignant neoplasm of colon	0	Exclude	Unsure whether to include secondary malignant
B575100	Secondary malignant neoplasm of rectum	0	Exclude	Unsure whether to include secondary malignant
B575z00	Secondary malig neop of large intestine or rectum NOS	0	Exclude	Unsure whether to include secondary malignant
B57y.00	Secondary malignant neoplasm of other digestive organ	0		
B57z.00	Secondary malig neop of respiratory or digestive system NOS	0		
B58..00	Secondary malignant neoplasm of other specified sites	0		
B58..11	Secondary carcinoma of other specified sites	0		
B58yz00	Secondary malignant neoplasm of other specified site NOS	0		
B58z.00	Secondary malignant neoplasm of other specified site NOS	0		
B59..00	Malignant neoplasm of unspecified site	0		
B590.00	Disseminated malignancy NOS	0		
B590.11	Carcinomatosis	0		
B591.00	Other malignant neoplasm NOS	0		
B592.00	Malignant neoplasms of independent (primary) multiple sites	0		
B593.00	Primary malignant neoplasm of unknown site	0		
B594.00	Secondary malignant neoplasm of unknown site	0		
B595.00	Malignant tumour of unknown origin	0		
B59z.00	Malignant neoplasm of unspecified site NOS	0		
B5y..00	Malignant neoplasm of other and unspecified site OS	0		
B5z..00	Malignant neoplasm of other and unspecified site NOS	0		
B713.00	Benign neoplasm of colon	0		Benign
B713.12	Benign neoplasm of ileocaecal valve	0		Benign
B713000	Benign neoplasm of hepatic flexure of colon	0		Benign
B713100	Benign neoplasm of transverse colon	0		Benign
B713200	Benign neoplasm of descending colon	0		Benign
B713300	Benign neoplasm of sigmoid colon	0		Benign
B713400	Benign neoplasm of caecum	0		Benign
B713500	Benign neoplasm of appendix	0		Benign

B713600	Benign neoplasm of ascending colon	0		Benign
B713700	Benign neoplasm of splenic flexure of colon	0		Benign
B713800	Benign neoplasm of colostomy site	0		Benign
B713900	Benign neoplasm of ileocaecal valve	0		Benign
B713z00	Benign neoplasm of colon NOS	0		Benign
B714.00	Benign neoplasm of rectum and anal canal	0		Benign
B714000	Benign neoplasm of rectosigmoid junction	0		Benign
B714100	Benign neoplasm of rectum	0		Benign
B714z00	Benign neoplasm of rectum or anal canal NOS	0		Benign
B718200	Benign neoplasm of retrocaecal tissue	0		Benign
B718300	Benign neoplasm of mesocolon	0		Benign
B718400	Benign neoplasm of mesorectum	0		Benign
B8...00	Carcinoma in situ	0		
B80..00	Carcinoma in situ of digestive organs	0		
B80..11	Ca-in-situ of G.I. tract	0		
B803.00	Carcinoma in situ of colon	0		TM - Polyp. Strictly speaking this is not a cancer. It is “in situ” if the malignant cells break do not through the basement membrane. It is a cancer if they do. Not all “in situ” become cancer – probably only a minority progress.
B803000	Carcinoma in situ of hepatic flexure of colon	0		Polyp
B803100	Carcinoma in situ of transverse colon	0		Polyp
B803200	Carcinoma in situ of descending colon	0		Polyp
B803300	Carcinoma in situ of sigmoid colon	0		Polyp
B803400	Carcinoma in situ of caecum	0		Polyp
B803500	Carcinoma in situ of appendix	0		Polyp
B803600	Carcinoma in situ of ascending colon	0		Polyp
B803700	Carcinoma in situ of splenic flexure of colon	0		Polyp
B803800	High grade dysplasia of colon	0		TM - Polyp. Again, not quite cancerous. The cells look more like malignant cells but because it is dysplasia it is confined within the basement membrane.
B803z00	Carcinoma in situ of colon NOS	0		Polyp
B804.00	Carcinoma in situ of rectum and rectosigmoid junction	0		Polyp
B804000	Carcinoma in situ of rectosigmoid junction	0		Polyp
B804100	Carcinoma in situ of rectum	0		Polyp
B804z00	Carcinoma in situ of rectum or rectosigmoid junction NOS	0		Polyp
B80z.00	Carcinoma in situ of other and unspecified digestive organs	0		Polyp
B80zz00	Carcinoma in situ of digestive	0		Polyp

	organs NOS			
B8yyz00	Carcinoma in situ of other specified site NOS	0		Polyp
B8z..00	Carcinoma in situ NOS	0		Polyp
B9...00	Neoplasms of uncertain behaviour	0		Polyp
B90..00	Neop uncertain behaviour of digestive and respiratory system	0		Not specific enough
B902.00	Neop of uncertain behaviour stomach, intestines and rectum	0		Not specific enough
B902400	Neoplasm of uncertain behaviour of colon	0		TM - I am not really sure what to say about this one. It implies it is not known to be malignant so I would not call it a cancer. It is not obvious that it refers to a polyp so I would not call it a polyp. So in my opinion neither cancer nor polyp
B902500	Neoplasm of uncertain behaviour of rectum	0		TM - I am not really sure what to say about this one. It implies it is not known to be malignant so I would not call it a cancer. It is not obvious that it refers to a polyp so I would not call it a polyp. So in my opinion neither cancer nor polyp
B902600	Neoplasm of uncertain or unknown behaviour of appendix	0		TM - I am not really sure what to say about this one. It implies it is not known to be malignant so I would not call it a cancer. It is not obvious that it refers to a polyp so I would not call it a polyp. So in my opinion neither cancer nor polyp
B902z00	Neop of uncertain behaviour stomach, intestine or rectum NOS	0		Not specific enough
B905.00	Neop of uncertain behaviour other and unspec digestive organ	0		
B93..00	Neop uncertain behaviour other and unspec sites and tissues	0		
B93y.00	Neoplasm of uncertain behaviour of other specified sites	0		
B93yz00	Neop of uncertain behaviour of other specified sites NOS	0		
B93z.00	Neop uncertain behaviour other unspec site and tissue NOS	0		
B9y..00	Neoplasm of uncertain behaviour otherwise specified	0		
B9z..00	Neoplasm of uncertain behaviour NOS	0		
BA...00	Unspecified nature neoplasm	0		
BA0..00	Neoplasm of unspecified nature	0		
BA00.00	Neoplasm of unspecified nature of digestive system	0		
BA0y.00	Neoplasm of unspecified nature of other specified sites	0		
BA0z.00	Neoplasm of unspecified nature NOS	0		
BAz..00	Neoplasm of unspecified nature NOS	0		
BB51000	[M]Adenocarcinoma in situ in villous adenoma	0		TM - Polyp. Within the polyp there are cancerous cells but they are confined by the basement membrane. Hence

				“pre-cancerous” and not cancer.
BB51100	[M]Adenocarcinoma in situ in tubulovillous adenoma	0		TM - Polyp. Within the polyp there are cancerous cells but they are confined by the basement membrane. Hence “pre-cancerous” and not cancer.
BB52000	[M]Adenocarcinoma in tubulovillous adenoma	1	Include	TM - almost certainly a bowel cancer
BB57.00	[M]Adenocarcinoma, intestinal type	0		Not specific enough
BB5L.00	[M]Adenomatous and adenocarcinomatous polyps	0		Polyp. - probably means there is more than one
BB5L000	[M]Adenomatous polyp NOS	0		
BB5L011	[M]Polypoid adenoma	0		Same as above synonym
BB5L100	[M]Adenocarcinoma in adenomatous polyp	1	Include	
BB5L200	[M]Adenocarcinoma in situ in adenomatous polyp	0		Polyp. Within the polyp there are cancerous cells but they are confined by the basement membrane. Hence “pre-cancerous” and not cancer.
BB5L300	[M]Adenocarcinoma in multiple adenomatous polyps	1	Include	
BB5Lz00	[M]Adenomatous or adenocarcinomatous polyp NOS	0		TM - Polyp
BB5M.00	[M]Tubular adenomas and adenocarcinomas	1	Include	
BB5M000	[M]Tubular adenoma NOS	0		TM - Polyp
BB5M100	[M]Tubular adenocarcinoma	1	Include	
BB5Mz00	[M]Tubular adenoma or adenocarcinoma NOS	1	Include	TM- cancer
BB5N.00	[M]Adenomatous and adenocarcinomatous polyps of colon	0		TM - polyp
BB5N.11	[M]Adenoma or or adenocarcinoma in polyposis coli	0		Higher risk not included
BB5N000	[M]Adenomatous polyposis coli	0		
BB5N011	[M]Adenomatosis NOS	0		
BB5N012	[M]Familial polyposis coli	0		Not including FAP as higher risk
BB5N100	[M]Adenocarcinoma in adenomatous polposis coli	0		
BB5N200	[M]Multiple adenomatous polyps	0		
BB5N211	[M]Multiple polyposis	0		
BB5Nz00	[M]Adenomatous or adenocarcinomatous polyps of the colon NOS	0		TM - polyp
BB5R200	[M]Carcinoid tumour, argentaffin, NOS	0		TM - exclude
BB5R211	[M]Argentaffinoma NOS	0		TM - exclude
BB5R300	[M]Carcinoid tumour, argentaffin, malignant	0		TM - Exclude
BB5R600	[M]Mucocarcinoid tumour, malignant	0		
BB5R611	[M]Goblet cell tumour	0		
BB5R700	[M]Composite carcinoid	0		
BB5R800	[M]Adenocarcinoid tumour	1	Include	Form of appendiceal carcinoid
BB5R900	[M]Neuroendocrine carcinoma	0		

BB5RA00	[M]Merkel cell carcinoma	0		
BB5Rz00	[M]Carcinoid tumours NOS	0		
BB5U.00	[M]Villous adenomas and adenocarcinomas	1	Include	
BB5U000	[M]Villous adenoma NOS	0		Type of Polyp
BB5U011	[M]Villous papilloma	0		Type of Polyp
BB5U100	[M]Adenocarcinoma in villous adenoma	1	Include	TM - Cancer
BB5U200	[M]Villous adenocarcinoma	1	Include	
BB5U300	[M]Tubulovillous adenoma	0		Type of polyp
BB5U311	[M]Papillotubular adenoma	0		Type of polyp
BB5U312	[M]Villoglandular adenoma	0		Type of polyp
BB5Uz00	[M]Villous adenoma or adenocarcinoma NOS	1	Include	TM - cancer
BB6..00	[M]Adnexal and skin appendage neoplasms	0		
BB60.00	[M]Skin appendage adenoma and carcinoma	0		
BB60000	[M]Skin appendage adenoma	0		
BB60100	[M]Skin appendage carcinoma	0		
BB60z00	[M]Skin appendage adenoma or carcinoma NOS	0		
BB6z.00	[M]Adnexal and skin appendage neoplasm NOS	0		
BB7..00	[M]Mucoepidermoid neoplasms	0		
BB70.00	[M]Mucoepidermoid tumour	0		
BB71.00	[M]Mucoepidermoid carcinoma	0		
BB7z.00	[M]Mucoepidermoid neoplasm NOS	0		
BB8..00	[M]Cystic, mucinous and serous neoplasms	0		
BB82000	[M]Mucinous adenoma	0		
BB82100	[M]Mucinous adenocarcinoma	0		
BB82111	[M]Colloid adenocarcinoma	0		
BB82112	[M]Gelatinous adenocarcinoma	0		
BB82113	[M]Mucoid adenocarcionoma	0		Not specific enough
BB82114	[M]Mucous adenocarcinoma	0		Not specific enough
BB82z00	[M]Mucinous adenoma or adenocarcinoma NOS	0		
BB83.00	[M]Pseudomyxoma peritonei	1	Include	Rare type of cancer that usually begins in your appendix
BB84.00	[M]Mucin-producing adenocarcinoma	0		Not specific enough
BB85.00	[M]Signet ring carcinoma	0		Not specific enough
BB85000	[M]Signet ring cell carcinoma	0		Not specific enough
BB85100	[M]Metastatic signet ring cell carcinoma	0		Not specific enough
BB85111	[M]Krukenberg tumour	0		Not specific enough
BB85z00	[M]Signet ring carcinoma NOS	0		Not specific enough
BB8z.00	[M]Cystic, mucinous or serous neoplasm NOS	0		Not specific enough
BBL..00	[M]Complex mixed and stromal neoplasms	0		Not specific enough
By...00	Neoplasms otherwise specified	0		

Byu..00	[X]Additional neoplasm classification terms	0		
Byu1.00	[X]Malignant neoplasm of digestive organs	0		
Byu1200	[X]Malignant neoplasm of intestinal tract, part unspecified	0		
Byu1300	[X]Malignant neoplsm/ill-defin sites within digestive system	0		
ByuC.00	[X]Malignant neoplasm of ill-defined, secondary and unspeci	0		
ByuC000	[X]Malignant neoplasm of other specified sites	0		
ByuC100	[X]Malignant neoplasm/overlap lesion/other+ill-defined sites	0		
ByuC400	[X]Secondary malignant neoplasm/oth+unspcfd digestive organs	0		
ByuC700	[X]Secondary malignant neoplasm of other specified sites	0		
ByuC800	[X]Malignant neoplasm without specification of site	0		
ByuE.00	[X]Malignant neoplasms/independent (primary) multiple sites	0		
ByuE000	[X]Malignant neoplasms/independent(primary)multiple sites	0		
ByuF.00	[X]In situ neoplasms	0		
ByuF000	[X]Carcinoma in situ/other+unspecified parts of intestine	0		
ByuF100	[X]Carcinoma in situ of other specified digestive organs	0		
ByuF200	[X]Carcinoma in situ of digestive organ, unspecified	0		
ByuFE00	[X]Carcinoma in situ of other specified sites	0		
ByuH.00	[X]Neoplasms of uncertain and unknown behaviour	0		
ByuH000	[X]Neoplasm/uncertain+unknown behaviour/oth digestive organs	0		
ByuH900	[X]Neoplasm/uncertain+unknown behavior/other specified sites	0		
ByuHA00	[X]Neoplasm of uncertain and unknown behaviour, unspecified	0		
Bz...00	Neoplasms NOS	0		
F011600	Meningitis due to adenovirus	0		
F011611	Adenovirus meningitis	0		
ZV10017	[V]Personal history of malignant neoplasm of rectum	0		Not a new diagnosis? TM- I agree, history of past cancer
ZV76400	[V]Screening for malignant neoplasm of colon or rectum	0		

Table A.2.2: Bowel Cancer Diagnosis Read Code review list with 1st and 2nd reviewer decisions. 1= include, 0=exclude for JC decision.

Appendix 3 – Drug Code List for Laxative Drugs

Drugcode (Laxative drug 04/2017)	genericname
50937978	Sodium dihydrogen phosphate anhydrous 340mg / Sodium bicarbonate 250mg suppositories
52289979	Macrogol compound oral powder sachets NPF sugar free
53912979	Macrogol compound oral powder sachets npf
54399979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
55041978	Macrogol compound half-strength oral powder sachets NPF sugar free
55042978	Macrogol compound half-strength oral powder sachets NPF sugar free
55044978	Macrogol compound oral powder sachets NPF sugar free
55530979	Macrogol compound half-strength oral powder sachets NPF sugar free
55687978	Macrogol compound half-strength oral powder sachets npf sugar free
55691978	Macrogol compound oral powder sachets NPF sugar free
60079979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
60080979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
60081979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
60321979	Lactulose 10g/15ml oral solution 15ml sachets sugar free
61925979	Macrogol compound oral liquid npf sugar free
62538979	Bisacodyl 7.5mg suppositories
63328979	Ispaghula husk oral powder sugar free
64079979	Magnesium sulfate powder
64080979	Magnesium sulfate powder
64082979	Magnesium sulfate powder
64919979	Macrogol compound oral powder sachets NPF sugar free
67744994	Sodium citrate compound 5ml enema
69800979	Sodium dihydrogen phosphate dihydrate 680mg / sodium bicarbonate 500mg suppositories
69801979	Sodium dihydrogen phosphate dihydrate 680mg / sodium bicarbonate 500mg suppositories
70855979	Generic senokot dual relief tablets
71251979	Lactulose 10g oral powder sachets
71568979	Sodium acid phosphate 700mg / potassium dihydrogen phosphate 305mg tablets
73183978	Glycerol liquid
73184978	Macrogol compound half-strength oral powder sachets npf sugar free
74430978	Macrogol compound oral powder sachets NPF sugar free
74431978	Macrogol compound oral powder sachets NPF sugar free
74432978	Macrogol compound oral powder sachets NPF sugar free
76191978	Macrogol compound oral powder sachets NPF sugar free
78367979	Fig 500microlitres/5ml / senna fruit 400microlitres/5ml oral solution
78536978	Macrogol compound oral powder sachets NPF sugar free
78815979	Glycerol 750microlitres/5ml / sucrose 1.7g/5ml oral solution
79487979	Methylcellulose 200mg/5ml oral solution
80868979	Bisacodyl 2.5mg/5ml oral suspension
80870979	Bisacodyl 5mg/5ml oral suspension
80950998	Lactulose 10g/15ml oral solution 15ml sachets sugar free

81322998	Macrogol compound oral liquid NPF sugar free
81324998	Macrogol oral solution
81715998	Sodium picosulfate with magnesium citrate powder for oral solution
81843998	Generic senokot comfort tablets
81847998	Glycerol 750microlitres/5ml / sucrose 1.7g/5ml oral solution
81870998	Glycerol 750microlitres/5ml / sucrose 1.7g/5ml oral solution
81919998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
81938998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
81959998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
81972998	Ispaghula husk 90% granules sugar free
82005998	Macrogol compound oral powder sachets npf sugar free
82013998	Sodium picosulfate 2.5mg capsules
82156998	Sodium citrate compound 5ml enema
82288998	Bisacodyl 10mg/30ml enema
82305998	Macrogol 4000 10g oral powder sachets sugar free
82306998	Macrogol compound oral powder sachets NPF sugar free
82314998	Glycerol 750microlitres/5ml oral solution sugar free
82325998	Glycerol 750microlitres/5ml oral solution sugar free
82445998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
82446998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
82491998	Ispaghula husk 3.4g/sachet sugar free powder
82493998	Ispaghula husk 3.4g/sachet sugar free powder
82641998	Glycerol 750microlitres/5ml oral solution sugar free
82642978	Sodium dihydrogen phosphate anhydrous 680mg / sodium bicarbonate 500mg suppositories
82692998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
82693998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
82701998	Macrogol compound oral powder sachets NPF sugar free
82703998	Macrogol compound oral powder sachets NPF sugar free
82726978	Glycerol 500microlitres/5ml / ipecacuanha liquid extract 9.8microlitres/5ml oral solution
82757978	Glycerol 750microlitres/5ml oral solution sugar free
82761978	Lubiprostone 24microgram capsules
82762978	Lubiprostone 24microgram capsules
83053998	Macrogol compound oral powder sachets npf sugar free
83112998	Macrogol 13.7g powder
83113998	Macrogol compound oral powder sachets npf
83177978	Macrogol compound oral powder sachets NPF sugar free
83203998	Macrogol compound oral powder sachets NPF sugar free
83841998	Generic dual lax extra strong tablets
83888998	Macrogol compound oral powder sachets npf
84124998	Docusate 100mg capsules
84262978	Macrogol compound oral powder sachets NPF sugar free
84308998	Generic lepicol powder
84309998	Ispaghula husk with lactobacillus and bifidobacteria oral powder
84329998	Glycerol 750microlitres/5ml oral solution sugar free

84658998	Glycerol 1.36g/5ml / glucose liquid 280mg/5ml oral solution
84659998	Glycerol 1.36g/5ml / glucose liquid 280mg/5ml oral solution
84793998	Macrogol compound oral powder sachets NPF sugar free
84804998	Generic senokot dual relief tablets
84900979	Magnesium sulfate powder
84920979	Methylcellulose powder
85071998	Generic Moviprep A oral powder 112g sachets
85288998	Sodium acid phosphate 15.6% oral solution (1mmol/ml)
85334998	Sodium dihydrogen phosphate 15.6% oral solution (1mmol/ml)
85537978	Glycerol 4g suppositories
85840998	Docusate 100mg capsules
85843998	Bisacodyl 5mg gastro-resistant tablets
86028979	Liquid paraffin light liquid
86207979	Glycerol liquid
86367979	Macrogol compound half-strength oral powder sachets NPF sugar free
86476979	Macrogol 4000 10g oral powder sachets sugar free
86509979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
86510979	Ispaghula husk 3.5g sugar free granules
86510998	Senna 7.5mg/5ml oral solution sugar free
86511998	Senna 7.5mg/5ml oral solution sugar free
86512998	Senna 7.5mg/5ml oral solution sugar free
86587998	Ispaghula husk 3.5g/sachet sugar free granules
86588998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
86589998	Ispaghula husk 3.5g/sachet sugar free granules
86890998	Cascara dry extract 130mg / senna leaf 32mg tablets
86891998	Cascara dry extract 130mg / senna leaf 32mg tablets
86955998	Ispaghula husk 90% granules
87141979	Sterculia 62% granules 7g sachets gluten free
87142979	Sterculia 62% granules gluten free
87143979	Sterculia 62% granules gluten free
87144979	Sterculia 62% granules gluten free
87146979	Sterculia 62% / Frangula 8% granules 7g sachets gluten free
87147979	Sterculia 62% / Frangula 8% granules 7g sachets gluten free
87148979	Sterculia 62% / Frangula 8% granules gluten free
87149979	Sterculia 62% / Frangula 8% granules 7g sachets gluten free
87217979	Senna fruit 12.4% / ispaghula 54.2% granules 4g sachets
87278998	Sodium dihydrogen phosphate dihydrate 1.69g / sodium bicarbonate 1.08g suppositories
87282998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
87283998	Ispaghula husk 3.5g/sachet sugar free granules
87624998	Glycerol 750microlitres/5ml / sucrose 1.7g/5ml oral solution
87625998	Glycerol 750microlitres/5ml / sucrose 1.7g/5ml oral solution
87626998	Glycerol 750microlitres/5ml oral solution sugar free
87650998	Macrogol compound half-strength oral powder sachets NPF sugar free
88071998	Macrogol 4000 10g powder

88168998	Senna 15mg tablets
88181998	Methylcellulose 4% solution
88409998	Lactulose 3.1-3.7g/5ml oral solution
88463998	Docusate 50mg/5ml oral solution sugar free
89059997	Ispaghula husk gluten-free sugar free granules
89059998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
89127998	Macrogol compound oral powder sachets NPF sugar free
89227998	Senna 15mg tablets
89230998	Macrogol compound half-strength oral powder sachets npf sugar free
89364998	Glycerol 4g suppositories
89374998	Glycerol 750mg/5ml oral solution
89380998	Ispaghula husk 3.5g/sachet powder
89381997	Senna 15mg tablets
89381998	Senna 15mg tablets
89383998	Senna 7.5mg tablets
89417998	Senna 15mg tablets
89550998	Phosphates formula b enema
89551998	Phosphates enema (Formula B) 128ml long tube
89552998	Phosphates enema (Formula B) 128ml standard tube
89617998	Ispaghula husk 3.5g sugar free granules
89619979	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
89644998	Magnesium sulfate powder
89795998	Sodium picosulfate 5mg/5ml oral solution sugar free
89886998	Phenolphthalein yellow 120mg / rhubarb 27.5mg tablets
89887998	Phenolphthalein yellow 120mg / rhubarb 27.5mg tablets
89906998	Bisacodyl 5mg gastro-resistant tablets
90122998	Senna 15mg tablets
90221998	Lactulose 3.1-3.7g/5ml oral solution
90275998	Sodium sulphate ppwder
90297979	Sodium citrate compound 5ml enema
90297998	Senna 7.5mg tablets
90298979	Sodium citrate compound 5ml enema
90347998	Liquid paraffin with magnesium hydroxide and sodium bicarbonate suspension
90348998	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
90459998	Aloin 38mg tablets
90460998	Aloin 38mg tablets
90494998	Dioctyl sulphosuccinate with phenolphthalein 100mg+60mg tablets
90642998	Macrogol compound half-strength oral powder sachets NPF sugar free
90671998	Sodium lauryl sulphoacetate with sodium citrate enema
90672998	Sodium citrate with sodium lauryl sulphoacetate enema
90673998	Sodium phosphate with sodium acid phosphate (10.8g with 24.4g)/45ml oral solution sugar free
90676998	Sodium dihydrogen phosphate dihydrate 542mg/ml / Disodium hydrogen phosphate dodecahydrate 240mg/ml oral solution sugar
90677998	Sodium citrate with sodium lauryl sulphoacetate enema
90678998	Sodium dihydrogen phosphate dihydrate 542mg/ml / Disodium hydrogen phosphate dodecahydrate

	240mg/ml oral solution sugar
90841998	Macrogol compound oral powder sachets NPF sugar free
90859998	Senna 12mg tablets
90860998	Senna 12mg pills
90901998	Macrogol 4000 10g oral powder sachets sugar free
91375998	Bisacodyl 5mg gastro-resistant tablets
91481997	Magnesium hydroxide oral suspension
91604998	Methylcellulose 450 liquid
91945998	Glycerol 750microlitres/5ml oral solution sugar free
92212998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
92241998	Senna 2.5mg/5ml oral solution
92363990	Macrogol compound oral powder sachets NPF sugar free
92566998	Ispaghula husk 3.4g/sachet powder
92570997	Ispaghula husk 6g/sachet sugar free powder
92570998	Ispaghula husk 3.4g/sachet powder
92606990	Phosphates enema (Formula B) 128ml long tube
92607990	Phosphates enema (Formula B) 128ml standard tube
92663998	Lactulose sachets
92664998	Lactulose 10g/sachet powder
92777998	Lactulose 3.1-3.7g/5ml oral solution
92791990	Sodium picosulfate 5mg/5ml oral solution sugar free
92839998	Sodium dihydrogen phosphate dihydrate 18.1% / Disodium hydrogen phosphate dodecahydrate 8% 133ml enema
92939998	Lactulose 3.1-3.7g/5ml oral solution
93004990	Docusate 100mg capsules
93049992	Sodium citrate compound 5ml enema
93125992	Ispaghula husk 3.5g/sachet sugar free granules
93295997	Magnesium hydroxide 415mg/5ml oral suspension sugar free
93295998	Magnesium hydroxide 300mg chewable tablets
93329998	Wheat fibre powder
93392992	Cascara eli
93443992	Ispaghula husk micronised + dextrose
93510998	Liquid paraffin liquid
93515998	Sterculia 62% / alverine 0.5% granules
93758998	Senna 2.5mg/5ml oral solution
93763992	Normacol x gra
93810998	Sodium dihydrogen phosphate anhydrous 1.936g effervescent tablets
93822998	Liquid paraffin & light liquid paraffin 7%+18% emulsion
93841992	Celevac liq
93924998	Methylcellulose 400mg tablets
93967998	Sodium citrate compound 5ml enema
93968998	Sodium citrate compound 5ml enema
93992992	Trifyba wheat husk pow
94120998	Magnesium sulphate enema
94121998	Magnesium sulphate enema

94132998	Arachis oil 130ml enema
94165998	Sodium acid phosphate & sodium bicarbonate 1.69g+1.08g suppositories
94202996	Magnesium hydroxide with aluminium hydroxide 300mg+600mg/5ml oral suspension
94203997	Magnesium hydroxide with aluminium hydroxide 300mg + 600mg tablet
94347998	Phenolphthalein with liquid paraffin liquid
94348998	Phenolphthalein with belladonna & ipecacuhana tablet
94351998	Magnesium hydroxide with oxetacaine and aluminium hydroxide 100mg+10mg+200mg/5ml oral suspension
94359998	Magnesium hydroxide with ambutonium and aluminium hydroxide sugar free oral suspension
94399998	Senna fruit 12.4% / Ispaghula 54.2% granules
94459996	Bisacodyl 2.74mg/ml rectal solution
94459997	Bisacodyl 10mg suppositories
94676990	Senna 7.5mg tablets
94700990	Glycerol liquid
94728992	Bisacodyl 10 mg tab
94746998	Senna fruit 12.4% / Ispaghula 54.2% granules
94775992	Methylcellulose granules
94820990	Magnesium sulfate powder
94915990	Liquid paraffin light liquid
94926992	Dioctyl sodium sulphosuccinate 20 mg tab
94976998	Docusate compound 5ml enema
94996998	Bisacodyl with dioctyl sodium sulphosuccinate tablets
94998998	Bisacodyl 5mg gastro-resistant tablets
95023992	Phosphates enema (Formula B) 128ml standard tube
95062992	Glycerol 4g suppositories
95066992	Glycerin & ichthammol sup
95067992	Glycerin 150 mg sup
95161992	Sterculia 62% granules gluten free
95252998	Sterculia 62% / Frangula 8% granules 7g sachets gluten free
95253997	Sterculia 80% granules
95253998	Sterculia 62% granules gluten free
95267992	Ispaghula husk 49% powder
95271998	Generic Picolax oral powder 16.1g sachets sugar free
95272996	Sodium picosulfate 2.5mg capsules
95272997	Sodium picosulfate powder for oral solution
95272998	Sodium picosulfate 5mg/5ml oral solution sugar free
95277992	Methylcellulose mouthwash sol
95295998	Sodium dihydrogen phosphate dihydrate 1.69g / sodium bicarbonate 1.08g suppositories
95296996	Phosphates enema (Formula B) 128ml long tube
95296997	Phosphates enema (Formula B) 128ml long tube
95296998	Phosphates enema (Formula B) 128ml standard tube
95297998	Sodium acid phosphate suppositories
95298996	Senna 15mg/5ml granules
95298997	Senna 7.5mg/5ml oral solution sugar free
95298998	Senna 7.5mg tablets

95336992	Sterculia 62% granules gluten free
95337992	Sterculia 62% granules gluten free
95344998	Ispaghula husk 49% powder
95345996	Ispaghula husk 6g oral powder sachets gluten free sugar free
95345997	Ispaghula husk 3.4g oral powder sachets gluten free sugar free
95345998	Ispaghula husk gluten-free sugar free 3.6g effervesant powder
95346996	Ispaghula husk gluten-free 3.4g powder
95346997	Ispaghula husk gluten-free sugar free granules
95346998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
95408992	Liquid paraffin & light liquid paraffin 7%+18% emulsion
95548998	Phenolphthalein with magnesium sulphate tablet
95549996	Phenolphthalein 120mg chewable tablet
95724992	Vegetable laxative tab
95726992	Ispaghula husk 66% granules
95731990	Senna 7.5mg tablets
95775992	Veracolate tab
95813992	Sodium picosulphate/magnesium cit sach
95857998	Methylcellulose-450 500mg tablets
95858996	Methylcellulose 900mg/10ml gel
95858997	Methylcellulose 64% granules
95858998	Methylcellulose 500mg tablets
95900992	Colocynth & jalap co (vegetable laxative tab
95923998	Ispaghula husk 3.5g / Mebeverine 135mg effervescent granules sachets sugar free
95948992	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
95951990	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
95966990	Senna 7.5mg tablets
96030996	Lactulose 3.1-3.7g/5ml oral solution
96030997	Lactulose 3.35g/5ml flavoured Oral solution
96030998	Lactulose 3.35g/5ml Oral solution
96055992	Phenolphthalein gum 97.2 mg
96172990	Lactulose 3.1-3.7g/5ml oral solution
96297990	Lactulose 3.1-3.7g/5ml oral solution
96319992	Magnesium sulphate/phenolphthalein 300 mg tab
96320996	Docusate 100mg capsules
96320997	Docusate 50mg/5ml oral solution sugar free
96320998	Docusate sodium 100mg tablets
96321998	Docusate sodium with sorbitol enema
96322998	Docusate sodium with bisacodyl tablets
96323997	Docusate 100mg capsules
96325996	Docusate 120mg/10g enema
96325997	Docusate 50mg/5ml oral solution sugar free
96325998	Docusate sodium 100mg tablets
96347990	Senna 7.5mg tablets
96368990	Lactulose 3.1-3.7g/5ml oral solution

96388992	Phenolphthalein 130 mg tab
96413990	Bisacodyl 5mg gastro-resistant tablets
96439992	Sennoside b 15 mg gra
96622990	Liquid paraffin liquid
96700992	Liq paraffin & phenolphthalen mix
96758990	Lactulose 3.1-3.7g/5ml oral solution
96816990	Lactulose 3.1-3.7g/5ml oral solution
96909992	Lactulose 3.1-3.7g/5ml oral solution
96980992	Bran (wheat) 2 gm tab
97064990	Liquid paraffin oral emulsion
97065990	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
97081990	Bisacodyl 10mg suppositories
97198992	Wheat husk extract (concentrated) pow
97205992	Cremaffin emu
97311992	Dioctyl sodium/sorbitol enema .1 % liq
97401997	Senna 15mg/5ml granules
97401998	Senna 7.5mg/5ml oral solution sugar free
97402998	Senna 7.5mg tablets
97408998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
97409998	Ispaghula husk 3.5g effervescent granules sachets gluten free sugar free
97444992	Fibrous grain extract/fibrous citrus ext 375 mg tab
97534992	Glycerol 4g suppositories
97543988	Glycerol 4g suppositories
97623992	Ispaghula husk 66 % gra
97629992	Ispaghula husk 90 % gra
97654998	Liquid paraffin light liquid
97655998	Docusate compound 5ml enema
97658990	Glycerol 1.36g/5ml / glucose liquid 280mg/5ml oral solution
97664992	Liq paraffin/light liq paraffin 7 % mix
97665992	Liq paraffin sterile liq
97666992	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
97667992	Liq paraffin & cascara emulsion emu
97668992	Liq paraffin emulsion mix
97691990	Liquid paraffin liquid
97710990	Senna 7.5mg tablets
97743992	Methylcellulose 450 mg/5ml liq
97744990	Lactulose 3.1-3.7g/5ml oral solution
97744992	Methylcellulose 900 mg mix
97750990	Glycerol 4g suppositories
97752998	Docusate 120mg/10g enema
97782990	Bisacodyl 5mg gastro-resistant tablets
97868992	Paraffin oil liq
97907990	Senna 7.5mg tablets
97945990	Glycerol 4g suppositories

98023992	Senna 7.5mg tablets
98044992	Phosphates enema (Formula B) 128ml long tube
98051998	Cascara tablets
98052998	Castor oil liquid
98053998	Figs compound liquid
98054998	Liquid paraffin oral emulsion
98062998	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
98095998	Senna 1mg/ml liquid
98099990	Liquid paraffin / Magnesium hydroxide oral emulsion sugar free
98127992	Sterculia 98 % gra
98153992	Syrp of figs liq
98159990	Glycerol 4g suppositories
98160990	Glycerol 4g suppositories
98215998	Methylcellulose granules
98243992	Vita fiber tab
98249998	Senna & ispaghula 12.4%+54.2% granules
98251997	Phosphates formula b enema
98251998	Phosphates enema (Formula B) 128ml standard tube
98252998	Arachis oil retention enema
98255998	Sterculia 62% granules gluten free
98256998	Sterculia 62% / Frangula 8% granules gluten free
98269998	Wheat fibre powder
98272998	Bisacodyl 2.74mg/ml rectal solution
98274998	Bisacodyl 10mg suppositories
98433998	Magnesium sulfate powder
98437996	Magnesium hydroxide 300mg chewable tablets
98437997	Magnesium hydroxide oral suspension
98437998	Magnesium hydroxide oral suspension
98453998	Magnesium sulphate with phenolphthalein tablet
98509990	Liquid paraffin liquid
98556998	Lactitol 10g oral powder sachets
98557998	Lactitol 10g oral powder sachets
98583990	Bisacodyl 10mg suppositories
98584989	Bisacodyl 5mg gastro-resistant tablets
98585989	Bisacodyl 10mg suppositories
98585990	Bisacodyl 5mg gastro-resistant tablets
98651990	Lactulose 3.1-3.7g/5ml oral solution
98692990	Magnesium hydroxide oral suspension
98822988	Glycerol 750microlitres/5ml oral solution sugar free
98839990	Glycerol liquid
98859992	Vitafibre
98889998	Sodium citrate compound 5ml enema
98918996	Glycerol 4g suppositories
99019998	Ispaghula husk 3.5g / Mebeverine 135mg effervescent granules sachets sugar free

99086990	Methylcellulose powder
99136988	Liquid paraffin oral emulsion
99136989	Liquid paraffin liquid
99136990	Liquid paraffin light liquid
99137989	Liquid paraffin liquid
99137990	Liquid paraffin oral emulsion
99138990	Liquid paraffin liquid
99170990	Lactulose 3.1-3.7g/5ml oral solution
99171990	Lactulose 3.1-3.7g/5ml oral solution
99188997	Ispaghula husk 3.6g/sachet powder
99188998	Ispaghula husk 3.4g/sachet sugar free powder
99193990	Glycerol liquid
99194988	Glycerol 4g suppositories
99195988	Glycerol 4g suppositories
99210998	Bran tabs
99224998	Sterculia 80% granules
99245998	Sodium dihydrogen phosphate anhydrous 1.936g effervescent tablets
99268990	Bisacodyl 10mg suppositories
99340998	Sterculia 62% / Frangula 8% granules 7g sachets gluten free
99394998	Magnesium hydroxide with oxetacaine and aluminium hydroxide 100mg+10mg+200mg/5ml oral suspension
99401979	Glycerol 4g suppositories
99416979	Lactulose 3.1-3.7g/5ml oral solution
99417979	Lactulose 3.1-3.7g/5ml oral solution
99424979	Sodium picosulfate 5mg/5ml oral solution sugar free
99425998	Sodium citrate compound 5ml enema
99426979	Senna 7.5mg tablets
99433979	Senna 7.5mg tablets
99457979	Bisacodyl 5mg gastro-resistant tablets
99462979	Senna fruit 12.4% / Ispaghula 54.2% granules
99472990	Liquid paraffin light liquid
99473989	Liquid paraffin oral emulsion
99473990	Liquid paraffin liquid
99489998	Sodium picosulfate 5mg/5ml liquid
99509998	Phenolphthalein with magnesium sulphate tablet
99524998	Ispaghula husk 90% granules
99528990	Magnesium hydroxide oral suspension
99535990	Lactulose 3.1-3.7g/5ml oral solution
99536990	Lactulose 3.1-3.7g/5ml oral solution
99537990	Lactulose 3.1-3.7g/5ml oral solution
99577990	Glycerol liquid
99578988	Glycerol 4g suppositories
99579988	Glycerol 4g suppositories
99601998	Macrogol npf 10g powder
99608998	Bran tabs

99717998	Lactulose 3.35g/5ml syrup
99721997	Sodium picosulfate 2.5mg capsules
99721998	Bisacodyl 5mg gastro-resistant tablets
99722998	Docusate sodium with bisacodyl tablets
99780989	Bisacodyl 5mg gastro-resistant tablets
99780990	Bisacodyl 10mg suppositories
99781989	Bisacodyl 10mg suppositories
99781990	Bisacodyl 5mg gastro-resistant tablets
99815998	Methylcellulose 900mg/10ml mixture
99845998	Methylcellulose 500mg tablets
99903992	Rhubarb & soda ammoniated mix
99954989	Sodium acid phosphate powder
99954990	Sodium dihydrogen phosphate dihydrate powder
99955997	Magnesium hydroxide with simeticone and aluminium hydroxide oral suspension
99984990	Magnesium sulfate powder

Table A.3.1: Final drug code list for laxatives.

Appendix 4 – Stata Code for Select Examples

```
//------------------------------------------------------------------------------
//*****CODE TO DERIVE BCSP FOBT SCREENING OUTCOMES FROM AHD DATA*****
//------------------------------------------------------------------------------

//Load in the data
use ahd.dta
//------------------------------------------------------------------------------
keep if regexm(medcode, "6867\.00|6866\.00|686A\.00|686B\.00|686C\.00|9Ow2\.00")
//------------------------------------------------------------------------------

merge m:1 medcode using /Volumes/JC/thin1509/stata/systemlookups/readcodes.dta,
keep(master match)
drop _merge

//Generate new variable and make it = to description
gen FOBToutcome=description
tab FOBToutcome
tab description

  description |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
                BCSP FOB test abnormal |         92        1.86        1.86
                  BCSP FOB test normal |      4,465       90.35       92.21
           BCSP FOB testing kit spoilt |          1        0.02       92.23
        BCSP FOB tst incmplt participt |         87        1.76       93.99
Bowel cancer screening programme: fae.. |        269        5.44       99.43
No response to bowel cancer screening.. |         28        0.57      100.00
--------------------------------------+-----------------------------------
                                 Total |      4,942      100.00

//------------------------------------------------------------------------------

//replace FOBToutcome = "BCSP FOB test normal" if ((medcode=="6866.00|68W2.00") &
(data4== "PTH005|P/N001"))
replace FOBToutcome = "BCSP FOB test normal" if medcode=="6866.00" & data4== "PTH005"
replace FOBToutcome = "BCSP FOB test normal" if medcode=="6866.00" & data4== "P/N001"

replace FOBToutcome = "BCSP FOB test abnormal" if medcode=="6866.00" & data4==
"P/N002"
replace FOBToutcome = "BCSP FOB test abnormal" if medcode=="6866.00" & data4==
"PTH010"

tab FOBToutcome

  FOBToutcome |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
                BCSP FOB test abnormal |         95        1.92        1.92
                  BCSP FOB test normal |      4,611       93.30       95.22
           BCSP FOB testing kit spoilt |          1        0.02       95.24
        BCSP FOB tst incmplt participt |         87        1.76       97.01
Bowel cancer screening programme: fae.. |        120        2.43       99.43
No response to bowel cancer screening.. |         28        0.57      100.00
--------------------------------------+-----------------------------------
                                 Total |      4,942      100.00
//------------------------------------------------------------------------------
//When there is a generic BCSP readcode and no other value recorded, this is not
helpful so drop it
drop if medcode=="6866.00" & !regexm(data4, "PTH005|P/N001|P/N002|PTH010")
//------------------------------------------------------------------------------
tab FOBToutcome

  FOBToutcome |      Freq.     Percent        Cum.
--------------------------------------+-----------------------------------
                BCSP FOB test abnormal |         95        1.97        1.97
                  BCSP FOB test normal |      4,611       95.62       97.59
           BCSP FOB testing kit spoilt |          1        0.02       97.62
        BCSP FOB tst incmplt participt |         87        1.80       99.42
No response to bowel cancer screening.. |         28        0.58      100.00
--------------------------------------+-----------------------------------
                                 Total |      4,822      100.00

//------------------------------------------------------------------------------
```

Table A.4.1: Stata code to derive BCSP FOBT screening outcomes from AHD data

```
//------------------------------------------------------------------------------
//*****BOWEL CANCER DIAGNOSIS*****
//------------------------------------------------------------------------------
use readcodes.dta
generate lcase=lower(description)
gen case = 0
//------------------------------------------------------------------------------

//Iterative search strategy

replace case=1 if regexm(lcase,"colo.*cancer")
replace case=1 if regexm(lcase,"colo") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"bowel") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"rect") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"caec") & regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"append") &
regexm(lcase,"neop|carcinoma|adeno|cancer")
replace case=1 if regexm(lcase,"git") & regexm(lcase,"neop|carcinoma|adeno|cancer")

browse if case ==1

//------------------------------------------------------------------------------
//Extract these into Excel document
//------------------------------------------------------------------------------

//Looking at additional stems identified of relevance from the above; codes beginning
//with B and 68 appear to be of relevance
//Look above and below for relevant terms
//Below gives the parent stems

browse if regexm(medcode, "^B.\.")

browse if regexm(medcode, "^B.\.*\.00")
browse if regexm(medcode, "^68.\.*\.00")
//------------------------------------------------------------------------------
//From the above, identified that the following stems may be of relevance
browse if regexm(medcode, "^B\..*|^B1.*|^B5.*|^B8.*|^B9.*|^BA.*|^BB.*|^By.*|^Bz.*") &
case ==0
browse if regexm(medcode, "^68\..*|^681.*|^686.*|^68P.*|^68Q.*|^68W.*|^68Z.*") & case
==0

//No additional codes identified from the above 68 parent stem
//------------------------------------------------------------------------------
//browse and extract those which are relevant into Excel document
//------------------------------------------------------------------------------

//Looking at BB(morphology in more detail as large number of results, to narrow down)
browse if regexm(medcode, "^BB.\.")

//Perhaps morphology is not relevant here, to confirm with Tom, otherwise indicate -
morphology could be relevant so include
//the child stems of relevance

browse if regexm(medcode,
"^BB.\.*|^BB0.*|^BB1.*|^BB2.*|^BB4.*|^BB5.*|^BB7.*|^BB8.*|^BBa.*|^BBB.*|^BBM.*|^BBm.*
|^BBT.*|^BBY.*|^BBy.*|^BBz.*") & case ==0

//------------------------------------------------------------------------------
//Reload final Excel back into Stata and sort by medcode before formatting in Excel

 import excel "/Users/jennifercooper/Desktop/Bham THIN/1.Creating lookups for
THIN/lookups_readcodes.xlsx", sheet("colorectalcancer (2)") cellrange(A3:F253)
firstrow clear

 sort medcode
//------------------------------------------------------------------------------
//END
```

Table A.4.2: Stata code used to derive a Read code list for Bowel Cancer Diagnosis

```
//------------------------------------------------------------------------------
//*****LAXATIVE DRUGS PROXY FOR CONSTIPATION*****
//------------------------------------------------------------------------------
use drugcodes.dta

//------------------------------------------------------------------------------
4 main types of laxative:

1.6.1 Bulk-forming laxatives
1.6.2 Stimulant laxatives
1.6.3 Faecal softeners
1.6.4 Osmotic laxatives

//Read text to determine which ones to include for acute constipation.
generate lcase=lower(genericname)
gen case = 0
//------------------------------------------------------------------------------
//Get an idea for keyword searches looking at Chapter from drug code dictionary
//as well as hardcopy BNF

browse if (regexm(bnfcode1, "^01.06")|regexm(bnfcode2, "^01.06")|regexm(bnfcode3,
"^01.06"))& case ==0
//------------------------------------------------------------------------------
//1. Search generic name for results based on drugs listed in the bnf.
//This allows you to identify drugs which have been mapped to different Chapters.
//Iterative search strategy

replace case=1 if
regexm(lcase,"ispaghula|fibrelief|fybogel|isogel|ispagel|regulan|methylcellulose|cele
vac|sterculia|normacol")
replace case=1 if
regexm(lcase,"bisacodyl|sodium.*picosulfate|dulcolax|pico.*liquid|pico.*perles|docusa
te.*sodium|dioctyl.*sodium.*sulphosuccinate|dioctyl|docusol|norgalax.*micro-
enema|glycerol|gylcerin|senna|sennoside|manevac|senokot|sodium.*picosulfate|dulcolax"
)
//We wont include Dantron as this is for terminally ill patients
replace case=1 if regexm(lcase,"arachis.*oil|liquid.*paraffin")
replace case=1 if regexm(lcase,
"lactulose|macrogol|polyethylene.*glycol|laxido|molaxole|movicol|norgine|magnesium.*s
alt|magnesium.*hydroxide|magnesium.*sulphate|phosphates.*rectal|carbalax|sodium.*acid
.*phosphate|sodium.*dihydrogen.*phosphate|fleet|casen.*fleet|phosphates.*enema|sodium
.*citrate|microlette.*micro.*enema|micralax|relaxit")

//------------------------------------------------------------------------------
//Put this stage in after investigations of tabulating the bnfcodes to see if we
//can remove any common ones that are not 01.06

//Get rid of those matched to Chapter 13, Chapter 7
//At this stage get rid of those which mention another Chapter as we re-add
//the drugs mapped to 01.06 below.

replace case=0 if regexm(bnfcode1, "^13\..*|^07\..*")
//------------------------------------------------------------------------------
//2. search the BNF Chapters (identified from keyword search as well as from hardcopy
book)
//for those where case=0

//We are interested in:
//1.6.1 Bulk-forming laxatives
//1.6.2 Stimulant laxatives
//1.6.3 Faecal softeners
//1.6.4 Osmotic laxatives

//Also interested in those drugs not mapped to a Chapter - 0 or 1

replace case=1 if (regexm(bnfcode1, "^01\.06\.00")|regexm(bnfcode2,
"^01\.06\.00")|regexm(bnfcode3, "^01\.06\.00"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.01")|regexm(bnfcode2,
"^01\.06\.01")|regexm(bnfcode3, "^01\.06\.01"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.02")|regexm(bnfcode2,
"^01\.06\.02")|regexm(bnfcode3, "^01\.06\.02"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.03")|regexm(bnfcode2,
"^01\.06\.03")|regexm(bnfcode3, "^01\.06\.03"))& case ==0
replace case=1 if (regexm(bnfcode1, "^01\.06\.04")|regexm(bnfcode2,
"^01\.06\.04")|regexm(bnfcode3, "^01\.06\.04"))& case ==0
```

```
//------------------------------------------------------------------------------

//exclude

//we do not want to include dantron as for terminally ill patients
replace case=0 if regexm(lcase, "dantron|co.*danthramer|co.*danthrusate")

//Scan through the types of drugs genericname
//we dont want creams, dressings etc
replace case=0 if regexm(lcase,
"dressing|poultice|eye|paediatric|injection|ear.*drop|cream|biscuit|syringe|ointment|
bath.*additive|bath.*oil|emollient|soap|shampoo")

//Can we remove any by formulation?
tab formulation
replace case=0 if regexm(formulation, "dressings|drops|paediatric|infant
suppositories")


//double check there are not any further ones that could be removed.

//------------------------------------------------------------------------------
browse if case==1
keep if case==1

//569 observations
//------------------------------------------------------------------------------

//Looking at which Chapters we could potentially remove

tab bnfcode1
tab bnfcode2
tab bnfcode3

//Go by bnfcode1 as that is the first Chapter the drug is mapped to?

//All combinations below:

preserve
        gen dummy=1
        collapse (count) dummy, by (bnfcode1 bnfcode2 bnfcode3)
        sort dummy
restore

//71 combinations, can we get rid of some which definitely do not map to the right
Chapter?
//13 seems to be a common Chapter for skin, we could remove this above by replacing
case with 0
//before we add all those Chapters that mention 01.06
//Also Chapter 7 obstetrics is a relatively common Chapter.
//------------------------------------------------------------------------------

//Exclude drugs which do not seem relevant (do not match drugs in the correct
Chapter)
//and do not have a bnf code which matches onto the correct Chapter.

//------------------------------------------------------------------------------
//ATC search as final check for drug ingredients


//------------------------------------------------------------------------------
//Final list for checking
keep if case==1

//------------------------------------------------------------------------------
```

Table A.4.3: Stata code used to derive a Drug code list for laxative drugs

Thesis Discussion

1.0 Summary of Findings

The studies in this thesis contribute to determining the value of risk adjusted colorectal cancer screening using the FIT through developing risk prediction models which can be used to guide referral decisions in screening.

The systematic review aimed to identify risk prediction models which combine the FIT result with risk predictors for colorectal cancer screening referral decisions and to determine whether they perform better than screening using the FIT alone. Before developing a model it is best practice to identify and build upon risk prediction models which have been previously developed. This review identified that there was some evidence to suggest that including additional predictors with the FIT result can improve model performance and test accuracy, relative to FIT only. Biomarkers (e.g. TIMP-1, sCD26 and CEA) and routinely available demographic factors (e.g. age, sex, BMI) gave improved performance metrics which suggested that improvement could be achieved using routine data alone without additional laboratory testing.

Routine data were used to develop the risk prediction models using both the BCSS and the anonymised primary care database THIN. The interconnectivity of GP records with other healthcare systems including screening is described in **Chapters 5 & 6**. Data for participants invited to the NHS BCSP are drawn from GP records and onto the NHS Spine. These connections can be exploited for further research.

This thesis uses three types of modelling strategy to develop risk prediction models: logistic regression, which is seen as the most common method used in the literature; artificial neural networks which is a flexible machine learning approach; and survival analysis in the form of Cox Regression to make use of the longitudinal nature of primary care records. Parametric models were also explored as an extension to Cox regression using the same predictors to determine whether model performance and fit was improved.

The predictors considered for risk prediction model development from the BCSS included demographic characteristics such as age, sex, IMD and previous screening history. Temperature and its effect on FIT positivity was also investigated by utilising open source data from the UK Met Office. Richer predictors could be obtained from primary care

records, e.g. symptoms, diagnoses of previous conditions, lab test results, anthropometrics and drug prescriptions.

Model building strategies used backwards elimination and integrated some form of internal validation to adjust for optimism and improve generalisability of the models. The outcomes investigated used colorectal cancer and advanced adenoma for the logistic regression and neural network models. For the Cox Regression model, both colorectal cancer and colorectal polyps were considered as a combined endpoint. The risk prediction models are presented as risk equations to ensure reproducibility and to enable external validation if required.

Model performance was assessed using discrimination (area under the ROC curve) as well as calibration (calibration plots and Hosmer-Lemeshow statistics). Discrimination improved from 0.63 with FIT only to 0.66 for the risk-adjusted logistic regression model (p=0.01). This was further improved when applying an artificial neural network model using the same predictors with an AUC of 0.69. A ROC test confirmed this difference was significantly different (p<0.001). There was also a corresponding increase in test accuracy when applying the model as a test. At a threshold of 160 µg/g, which is the anticipated NHS BCSP cut point, the ANN had a sensitivity of 35.15% and a specificity of 85.57% compared to a sensitivity of 33.15% and specificity of 84.69% for the equivalent logistic regression model. The FIT only as a comparator had a sensitivity of 30.78% at the same specificity.

Two risk prediction models were developed in **Chapter 5** to help identify predictors which could be considered for inclusion in a future risk adjusted screening model. The first model combined the FOBT result (both positive and negative) with other predictors available from THIN to determine individual risk. The second was developed to investigate if the additional information from the electronic GP record could be used to make better screening referral decisions for those with negative FOBT results only. Optimism adjusted performance statistics for the model combining FOBT included a C-statistic of 0.850, c-slope of 0.991, D statistic 2.298 and R^2 of 0.558. The model for negative FOBTs only had a C-statistic of 0.650, c-slope of 0.944, D statistic 0.836 and R^2 of 0.144.

Parametric survival models were also investigated as an extension to these two models. The generalised gamma model had the best fit based on the AIC, cumulative hazard plots, Kaplan Meier function plots and Cox-Snell residuals for a sample population with both negative and positive FOBT results. The discrimination for the generalised gamma model

was very similar (0.859 (95% CI: 0.845, 0.872)) to the equivalent Cox Regression model (0.854 (95% CI: 0.841, 0.868)). Calibration was slightly better for the generalised gamma model but still comparable to the Cox regression model. A Wald test for the hypothesis of the kappa ancillary parameter of the gamma model being equal to 1 was significant suggesting a potentially good fit for the Weibull model also (C-statistic 0.854 (95% CI: 0.841, 0.868)). For the sample population with negative FOBT results only, the Gompertz model provided the best fit when comparing the AIC and residual plots of other parametric models. With regards to model performance, the discrimination of the Gompertz model was the same as the equivalent Cox regression (C-statistic 0.658 (0.633, 0.683)) with a similar calibration. The covariate effect estimates also showed very similar results between the Gompertz model and the Cox model. The choice of the most appropriate model depends on several factors including practical application, model fit, out of sample (external) validation, how well the model follows the baseline hazard and whether assumptions of the model are met. A Cox Regression model was used to estimate predicted probabilities in this instance as similar studies have used this approach aiding comparability of results[1 2] and due to the flexibility of the parameterisation of the baseline hazard.[3 4]

In order to derive valid data for the analyses in **Chapter 5** using the THIN database, the methods behind data extraction and for improving data validity are described in **Chapter 6**. The AEB date identified the point at which GP practices started to receive electronic notifications from the BCSP and was used as a layer of data quality assurance and to define a screening cohort for analysis. Data are stored in four main files in THIN using clinical coding systems: Patient File, Medical File, Therapy File and Additional Health Data File. The methods used to extract lab test results and other predictors from the AHD file were described. The methods used to develop Read Code lists to extract information on symptoms and diagnoses were described. Drug code list development to extract information relating to prescriptions was also detailed. The reporting of this chapter ties into the growing interest in electronic health records and the need for reproducibility and transparency in the methods used to analyse them.

2.0 Summary of Chapters

Chapter two reports a systematic review which aimed to identify risk prediction models which combine the FIT result with risk indicators for colorectal cancer screening referral decisions and to determine whether they perform better than regular screening using the FIT alone. Eight studies were included from reviewing 54 full text articles. Discrimination ranged from 0.676-0.960 for risk adjusted FIT (reported in 6/8 studies) and 0.683-0.902 for FIT only (reported in 4/8 studies). Calibration using the Hosmer-Lemeshow statistic ranged from 0.276-0.940 for risk adjusted FIT (reported in 3/8 studies); and calibration plots were presented in just one study. Where test accuracy measures were included (4/8 studies), sensitivity ranged from 21.9% to 88% for risk adjusted FIT at a range of set specificities from 90-97.7%. FIT only sensitivity ranged from 19.7% to 82% at the same specificities. Although this could not be tested formally in a meta-analysis, there was evidence to suggest both model performance and test accuracy improved by combining the FIT with other risk predictors. Both the integration of lab results/biomarkers and routinely available demographic factors gave improved performance metrics. Age and sex were the most consistently included predictors, high performing models also included lifestyle information such as smoking and alcohol consumption as well as family history of cancer. Lab test results associated with greater model performance included TIMP-1, CEA, sCD26 and calgranulin B. Further evidence is required to confirm which biomarkers and other predictors should be included. This suggested that improvement in discrimination and test accuracy could be achieved using routine data alone without additional laboratory testing. None of the models could be considered ready to apply in practice and most suffered from methodological issues with statistical analysis which was rated as high risk of bias for all but one of the identified studies. Seven studies used a form of logistic regression and only one used survival analysis. There were no studies in this review which assessed machine learning approaches, which have been shown to have similar or superior performance to more conventional regression techniques.

Chapter three developed a risk prediction model combining routinely available predictors from the BCSS with the FIT to determine whether model performance and test accuracy were improved in an average risk screening population, compared to FIT alone. Model development used logistic regression and backwards elimination with cross validation**.** The final model included, log of the FIT result, age, sex and screening history. Discrimination improved from 0.628 for FIT only to 0.659 for the risk-adjusted model. The sensitivity

improved from 30.78% to 33.15% at similar specificity using a threshold of 160 µg/g, which is an anticipated threshold for the NHS BCSP. The risk-adjusted screening algorithm detected 13 additional advanced adenomas compared to FIT only. The use of routine data has a distinct advantage as no additional data needed to be collected from participants. This risk model mainly improved detection in men and would need further investigation if applied in practice. Positive FIT results occur more frequently in men compared to women[5] without risk adjustment. The BCSS should be explored for further risk indicators in the future, particularly relating to previous screening results. Machine learning algorithms were identified as a potential avenue to explore in **Chapter 2** and have the potential to improve model performance and test accuracy further without additional data collection.

Chapter four develops a feed forward ANN using the same routine predictors investigated in **Chapter 3** and cross-validation used to aid comparison. There is evidence to suggest that ANN models outperform logistic regression in certain scenarios and it has been suggested that both methods are investigated in a complementary manner.[6-9] The final network had 5 input nodes, 3 hidden layer nodes and 1 output node with a weight decay of 0.01 (cross validated deviance was 2077.7). The network was pruned by removing 4 connections leaving the model with 18 connection weights, helping to improve the generalisability of the model. The AUC for the ANN was 0.69 compared with 0.66 for the logistic regression model, this was statistically significant ($p<0.001$). Calibration was good as indicated by the Hosmer-Lemeshow statistic (0.892), giving a similar result to the logistic regression model (0.898). In terms of test accuracy, the ANN had a higher sensitivity compared to the logistic regression model (35.15%). Eleven additional advanced adenomas were detected using the ANN over the logistic regression model. Model development is fully reported and the full equation given which is an often cited disadvantage of studies using machine learning approaches.[10] With the shift to larger and more complex electronic health data, machine-learning algorithms may be better placed to deal with larger amounts of data and non-linear predictors when compared with conventional models such as logistic regression. Another approach to improving model performance is to consider a richer set of predictors. Model performance metrics including Nagelkerke's R^2, AUC and the deviance suggested that the prediction of cancer/advanced adenomas is not fully captured by the predictors used in both the logistic regression and ANN models. The BCSS receives data for its participants from the NHS Spine which houses demographic information for those aged 60-74 drawn from GP records. There is capacity to draw further information from the NHS Spine or from GP records to improve screening referral decisions. EHRs from primary care

have a richer level of data than that available on the BCSS and may add a further dimension to risk prediction models.

Chapter five investigated the use of electronic primary care databases to improve colorectal cancer screening referral decisions. Potential predictors from the database which may enhance a future risk adjusted model were investigated along with the data completeness of these predictors. The THIN database of anonymised GP records was used to define a screening population by identifying practices which receive electronic BCSP notifications in England and for participants aged 60-74. The positivity of the FOBT in this cohort was similar to that reported in the literature (2.18%). Data were generally well recorded for reported symptoms (100%), smoking status had 99.44% completeness and alcohol consumption 78% completeness. The least complete factors included lab results (platelet count, MCV, and haemoglobin at around 45%, and ferritin at 8.59%). Univariable analysis using Cox Regression was used to estimate hazard ratios (HR) for >30 key clinical features of colorectal cancer driven from the literature. Screening based factors had the strongest association with colorectal cancer/polyps. Previous positive FOBT results had a HR of 5.03 (CI: 4.18-6.05) and previous polyps diagnosed before the latest FOBT result had a HR of 3.18 (CI: 2.77-3.66). The Cox Regression model which combined the FOBT result (n=98,303; 1197 colorectal cancer/polyps) included 13 predictors and 2 interactions including; MCV, various symptoms/diagnoses and whether previous polyps had been diagnosed. The optimism adjusted performance metrics gave a C-statistic of 0.850, c-slope of 0.991, D statistic 2.298 and R^2 of 0.558. Parametric models were investigated as an extension with the generalised gamma model providing marginally better model performance with a C-statistic of 0.859 versus 0.854 (apparent performance) for the equivalent Cox Regression model and slightly better calibration as reflected in the calibration plots.

Since the guaiac FOBT sensitivity is around 50%, analysis was then repeated for a population with negative FOBT results only to determine whether additional factors could be used for screening referral decisions despite a negative test result. The model investigating negative results only (n = 95,792; 587 colorectal cancer/polyps) included a similar pattern of variables. Performance metrics included a C-statistic of 0.650, C-Slope of 0.944, D statistic 0.836 and R^2 of 0.144. The parametric model with the best fit to the data was the Gompertz model. This model had the same discrimination as the equivalent Cox regression (C-statistic 0.658) with very similar calibration plots. This study has shown that there are several clinical predictors available from GP databases which are associated with

colorectal cancer and polyps for an average risk screening population. Furthermore, this research has identified predictors which could be considered for inclusion in a future risk adjusted screening model. The prediction models estimate an individual's absolute risk of colorectal cancer and these methods could be used in future models to identify participants at highest risk for screening referral decisions. Additional data could be drawn from primary care onto the BCSS using the NHS Spine to contribute to a referral algorithm.

Chapter six describes the methodology used to extract valid data from THIN for the analysis in **Chapter 5**. A method was developed to define acceptable periods of BCSP notifications for practices receiving electronic results – the acceptable electronic BCSP (AEB) date. The frequency of bowel cancer screening notifications received per month by the number of patients registered in a practice aged 60-74 was determined for each practice in THIN. An expected rate for each practice was also generated, based on a 50% uptake rate and adults being invited biennially. Line graphs, with the expected monthly rate of electronic notifications and the actual monthly rate by practice were visually examined for the start date by two reviewers. There was 97.8% agreement for practices included in this assessment. The agreed AEB start dates for these practices were used in **Chapter 5** to define the population and data used for analysis. AHD strategies for bowel cancer screening notifications identified a series of rules which could be used to extract BCSP FOBT screening outcomes for all patients in the THIN database. The AHD strategy for haemoglobin concentration identified a reference distribution using pathology lab data and transformed values where appropriate so that the results used the same units and were recorded in the same way for analysis. Results which were outside the range of the reference distribution were excluded from extraction. Read code list generation resulted in a list of 42 codes used for bowel cancer diagnosis after being subject to a double reviewing process. Drug code list generation for laxatives resulted in a list of 450 codes. The AEB date can be used in future studies as a layer of data quality assurance to ensure results used for analysis are ones which have been electronically received to the AHD records. There is a growing requirement to ensure Read code/Drug code lists as well as more recently their methods are transparent and available for future research studies. The methods for producing Read code lists, drug code lists and AHD variables are fully reported.

3.0 Original Contributions

This research adds to the growing body of methodological and empirical research for risk prediction models, screening and EHRs.

The systematic review crosses both diagnostic accuracy and risk prediction model paradigms by assessing the model as a test and including both PROBAST and QUADAS-2 for quality appraisal. This approach is also carried forward into **Chapters 3 & 4** when developing a logistic regression and neural network model; test accuracy (sensitivity and specificity) was assessed by setting the recall rate the same (this generates the same specificity) and analyzing the corresponding effect on sensitivity.

As far as can be identified this is the first instance of a risk adjusted approach combining the FIT in a prediction model being investigated in the UK. **Chapters 3 & 4** developed the model using screening data available directly from the BCSS. Using data routinely available from the screening programme makes it easier to implement such a model into clinical practice.

In addition, machine learning methods (artificial neural networks) have not been previously considered for developing a model which combines FIT for screening decisions. One of the main criticisms identified by the TRIPOD guidelines for machine learning algorithms is a lack of transparency in the methods adopted for model development. The model building process was fully described and the risk equation provided with the study to aid reproducibility and transparency.

A novel method of utilizing a primary care database for a screening population was developed and applied in **Chapters 5 & 6**. The AEB date allows a researcher to identify when electronic BCSP notifications started to be received by each GP practice. This is a mark of quality assurance of the data since before this time paper records would have been used which are often biased towards abnormal results and may be incomplete. This AEB date was used to help define a screening cohort from the THIN database for developing a risk prediction model.

The methodology for extracting Additional Health Data (AHD) for particular variables is, as far as can be identified, previously unreported. This is a more complex data structure to use compared to the Medical File and Therapy data. The methods for extracting FOBT screening outcomes as well as for haemoglobin results are fully reported along with the corresponding Stata commands to improve transparency in the area of EHRs. This has been recommended in recent guidelines and publications.[11 12]

The results reported in this thesis are in themselves an original contribution to the literature. Risk adjusted screening has been shown in **Chapters 2, 3 & 4** to perform better than using the screening test alone. Both model performance and test accuracy were improved using this approach and an increase in the advanced adenoma/colorectal cancer detection rate was observed. Further research is required once this model is updated and externally validated to confirm these results in the form of an impact analysis or randomised controlled trial of risk adjusted screening versus regular screening.

Further to this, a risk adjusted approach using logistic regression performs better than FIT alone, but an artificial neural network was shown to have even greater performance. The neural network can provide the absolute risk for each individual like with a logistic regression model. In addition, with the shift to larger and more complex electronic health data, machine-learning algorithms may be better placed to deal with larger amounts of data and non-linear predictors when compared with conventional methods.

The risk adjusted logistic regression model compared to FIT only showed a greater detection rate in men versus women. The neural network on the other hand helped to equalise this disparity. Just as when assessing new health technologies, which might change the spectrum or profile of diagnosed disease (as well as the corresponding balance of benefits and harms), risk prediction models need to be assessed in a similar way since in this instance the proportion of cancers being identified was greater in men.

4.0 Practical Implications

Prediction model studies must follow the complete pathway from model development, to external validation and finally model impact before they are considered for implementation in practice. This research has identified key predictors for use in a screening referral algorithm from the BCSS. There is capacity on this information system to include further predictors in a risk score. There are further BCSS variables which warrant investigation particularly relating to screening history. A risk adjusted approach using the routinely available predictors from BCSS has been discussed by key stakeholders in the BCSP and it has been suggested that additional research is carried out to identify further predictors from the BCSS and to externally validate an updated model.

Each of the models developed in this thesis gave the full model equation for external validation and absolute risk probabilities for screening referral by setting an appropriate risk threshold. This approach will allow those at higher risk to be referred onto colonoscopy, whilst those at lower risk are placed back into the screening pool for continued surveillance. This approach helps to maximise benefits, minimise harms and make the most effective use of a limited colonoscopy resource. Absolute risk probabilities can also be used for risk communication to improve screening uptake and informed decision making.[13 14] A Nomogram such as the one presented in **Chapter 5,** can be used for such a purpose.

There are several barriers which have hindered the widespread application of machine learning approaches such as ANN's in practice.[15-22] Reasons include the 'black box' nature of the models which are more difficult to interpret,[6 23] computational transportability for external validation or implementation,[16] clinician trust or acceptability, the lack of methods to update machine learning algorithms when applied in a new setting or for recalibration[19 22] and the lack of reporting the model equation or providing 'software objects'[16] in publications. Efforts were therefore made within this thesis to provide the full risk equation for external validation and to enable the algorithm to be applied in computers with minimal software requirements enhancing computational transportability. In addition, methods to aid the interpretability of the model were presented including Garson's algorithm to show relative importance of variables, neural network structure plots, patient profiles to show which risk factors may contribute to higher risk probabilities and the

predictiveness curve to assess the fit of the model as well as the clinical utility. ANN model development is described in detail in **Chapter 4** and the risk equation gives an absolute risk probability like with the logistic regression model.

The logistic regression model developed in **Chapter 3** detects more cancers/advanced adenomas in males compared to females but the neural network helps to level out this difference by increasing the number of high risk adenomas detected in women and halving the number of false positive results for women. This difference seen between the sexes warrants further investigation based on screening programme aims.

The analyses reported in **Chapter 5** use a GP database to identify predictors which could be considered for inclusion in a future risk adjusted screening model for screening referral decisions. Screening history variables had a strong association with the diagnosis of colorectal cancer/polyps. Predictors which retained significance in the multivariable models and which could be considered in future risk based models included: MCV result, smoking status, family history of gastrointestinal cancer, age, sex, abdominal pain/antispasmodic prescription, diarrhoea, flatulence and change in bowel habit. The interconnections between GP records and the BCSS can be exploited further by drawing off more information onto the NHS Spine. Further to this, other studies have shown the merit of using blood test results combined with screening tests,[24 25] and the MCV was retained in both prediction models in this research. Routine blood test results for those in the screening age range could be implemented in the future. For example, the NHS Health Check is offered to individuals aged 40-74 and this could include routine blood tests.

The choice of the most appropriate model to use in practice depends on a multitude of factors, not just model performance. Determining the most appropriate model to use whether it be a semi-parametric Cox regression model, parametric model, logistic regression or neural network will depend firstly on the nature of the data and underlying assumptions and then on model performance parameters. External validation and impact studies will then provide the necessary evidence on out of sample performance and patient outcomes. The parametric models in **Chapter 5** for instance were found to have a similar performance to the Cox Regression models. For parametric models, the hazard is assumed to follow a specific statistical distribution,[4] whereas the Cox model has more flexibility with no restrictions on the shape of the hazard. However, if parametric models fit the baseline

hazard more closely, then more accurate coefficient estimates can be derived (smaller standard errors) and they offer more with post-estimation. The choice of parametric model should also be based on whether the shape of the hazard follows the appropriate statistical distribution of the model.[26] Further investigations therefore of the external performance of these models would need to be investigated with a consideration of how they would be applied. For instance, considering whether being able to derive risks for multiple time points is a useful application in this screening setting.

From an external validation perspective, machine learning algorithms are often harder to assess if based on 'software objects' or if the model equation is not fully reported. For complex random forest models for example, equations could span many pages. The full equation for the neural network was fully reported in **Chapter 4** to allow external validation and to allow implementation into statistical software (such as the BCSS). Issues may be encountered if the model requires recalibration/updating when testing in a different setting as these methods are not fully developed for this type of algorithm.[16] The methods for externally validating a logistic regression model are however fully developed and updating the model only requires a simple change to the intercept. More recently, guidance for external validation of a Cox Regression model has also been published,[27 28] although out of sample validation requires interpolation or extrapolation whereas parametric models can estimate predictions at a number of time points[29].

A model may perform better in terms of discrimination and calibration, like with the neural network model versus the logistic regression model, but usability (including presentation of the model) and how the model would be applied also plays a part on whether implemented in practice. Neural networks as discussed above are more difficult to interpret and are therefore not 'trusted' as much as tried and tested medical statistical approaches. Acceptability of these different risk adjusted approaches to both clinicians and patients would also need to be investigated before implementing in practice. Acceptability is considered in several of the National Screening Committee criteria for deciding whether to implement or change a component of a screening programme (Criteria 6 and 12 for example). Other important elements which would need to be considered are the presentation of the risk information to patients, GPs or Specialist Screening Practitioners (SSPs). Currently the BCSP are planning on reporting either a positive or negative result to a patient based on a FIT threshold (not the underlying numerical result which relates to risk).

For informed decision making, knowing your level of risk may cause worry if just under the probability threshold chosen, or may help to push an individual to go for colonoscopy if they are at substantially higher risk, or even cause an individual to pursue lifestyle changes. Reporting and presenting the level of risk with one of these statistical models would therefore require further investigation.

The methods used for defining the AEB date and the resultant dates extracted for each practice can be used in future studies using BCSP data. The AEB date can be used to provide quality assurance to the data as well as to help define a screening cohort for analysis from data which is principally used for primary care based studies. Similar methods can be applied on other EHRs and for screening programmes in different regions. The methods reported in **Chapter 6** for AHD variable extraction, Read code list and drug code list development can be used or amended for other studies using primary care databases.

5.0 Future Research

Updating Models and external validation studies

The risk based models developed in **Chapters 3 & 4** can be refined and updated by utilising additional risk predictors available from the BCSS as well as using follow up information once the FIT is rolled out in 2018. This refined model could then be assessed in a further dataset using FIT for external validation (for instance, Scotland have implemented the FIT along with the Isle of Man).

The impact of the risk prediction model could be investigated by recalling an individual if either the FIT result alone or the risk based model suggests cancer could be detected at colonoscopy and assessing the corresponding diagnostic accuracy and patient outcomes. The systematic review identified that there were no prediction model impact studies in this area combining FIT with other factors. Based on the results reported in **Chapters 3 & 4**, this approach would result in an estimated additional 109 colonoscopies out of 40,000 people invited.

The BCSS has an inbuilt function of using 1/n data for screening participants. For the pilot, 1 out of every 28 invitations was assigned a FIT; a similar approach here could be used to

assess risk adjusted screening or to assign a range of thresholds so data can be retained for future analysis.

Many risk prediction models are developed but fewer undergo external validation and even fewer have an impact analysis.[27 30-32] Ideally, models should be externally validated by a separate research team as a gold standard approach to model development and validation. The models in this research, although they improve discrimination have a less than perfect performance (area under the curve range 0.66-0.69 and for Cox Regression from 0.65-0.85). Externally validated models which have less than perfect performance require an impact analysis to determine if a risk adjusted approach is better than usual care.[33] An intermediate step could apply decision modelling techniques to assess the potential consequences of the model.[33] Future research in this area should therefore consider the whole prediction model pathway to ensure models with high predictive performance and which show patient benefit are applied in clinical practice.

External validation of Cox models compared to logistic regression is not frequently cited in the literature due to higher complexity associated with the baseline hazard function. Recent guidance has however recently been published in this area.[28]

Risk Based Screening Intervals

Timing or the 'intended moment of using the model' is an important consideration of the CHARMS checklist.[34] A risk based approach could be implemented at various points along the colorectal cancer screening pathway. This research focused on a diagnostic risk prediction model at the time of the screening test in order to assist with identifying those at highest risk for referral. The PROGRESS research group argue that assessing future outcome of risk (prognosis) may be a better approach than assessing diagnosis.[35] They state that diagnosis is a dichotomy at a single point in time and clinicians now have greater access to continuous measures of risk.[35 36] Based on these arguments, future research could assess a baseline FIT result (along with other factors) to determine personalised screening intervals (or for post-screening surveillance). People at increased risk of colorectal cancer could receive more frequent screens whereas those at lower risk could receive fewer screens. This approach could make screening programmes more cost effective. Studies have shown for instance that the baseline FIT concentration is a predictor of incident

colorectal neoplasia and is also related to detection of colorectal neoplasia in the next screening round.[37 38]

Assessing Diagnostic Accuracy of a Model

Assessing the diagnostic accuracy of the risk prediction model used as a screening test, depends on the timing of the model and the subsequent potential role of the test. Three roles have been defined for a new test: replacement, triage and add-on, and a variety of study designs can be used for comparing a new test with an existing test.[39] The risk prediction model combining the FIT could be considered as a replacement test due to its higher accuracy. Once a model has been externally validated it can be assessed against the FIT only in an RCT or a paired study whereby the patients are tested with the model, the screening test and the reference standard. Alternatively, a model can be used as a triage before the screening test to preselect a population for further testing; studies have implemented questionnaires for this approach.[40] Finally, a prediction model can be used as an-add on after the screening test to reduce the number of false positives. Biomarkers may have this role in the future.[33]

Prediction model methods

The models developed using survival analysis applied Cox Regression as a modelling technique. There are more flexible parametric survival models which exist including the Royston-Parmar model.[41] **Chapter 5** does extend the Cox Regression model to other parametric models and compares the model fit, but the Royston-Parmar models offer unique advantages. For instance, they fit a restricted cubic spline allowing greater flexibility when modelling the baseline hazard function, can incorporate time dependent effects and estimate hazard rates at all time points. Future studies should consider comparing this model to other approaches and following the guidance in terms of reporting, as discussed by Ng *et al*.[42]

For studies using longitudinal data with repeatedly measured variables, there could be different approaches to incorporate these changes in a prediction model to improve individual risk prediction. For instance, the change in the FIT concentration over screening rounds could be a strong risk indicator for colorectal cancer. Furthermore, blood test

results or biomarker levels could change over time which may reflect an underlying disease. More guidance and research is required in this area for the best way to model these changes. A recent article published in the *Diagnostic and Prognostic Research* journal applied 6 different methods to model a repeatedly measured predictor; all measurements, a single best measurement, summary measurement (mean or maximum), the change between subsequent measurements, conditional measurements and growth curve parameters.[43] The latter was the most flexible and retained predictive quality. Approaches such as these should be considered in future studies to improve individual risk prediction performance.

Although the TRIPOD statement gives recommendations for the reporting of studies developing validating or updating a prediction model, the focus is on regression techniques. Some of the principles are equally valid to machine learning algorithms but further more specific items are required when reporting such studies. A common theme identified in the literature pertaining to ANNs for instance is that the model building procedure is often not adequately reported.[7] Model building approaches are also less formalised compared to logistic regression so each study tends to have a different approach. In order to produce well performing models which are reproducible, further guidelines on building neural networks and other machine learning algorithms should be developed to improve both external validation and confidence in a clinical setting.

Finally, future research could include the consideration of dynamic risk prediction models which continually learn or update based on new information.[44-46] The underlying prevalence of disease may change, or more information obtained on a risk cohort which would require an update of the original model coefficients.

Electronic Health Records

Lifestyle factors, lab parameters, symptoms and other conditions have been identified as predictors, which help to explain an individual's level of risk for colorectal cancer in **Chapter 5** and could be included in a future risk adjusted screening algorithm. The importance of such factors could be considered by offering routine blood tests as part of NHS Health Checks or drawing further information from GP records onto the screening database for use in a pre-selection algorithm. The interconnections between the IT systems

used for primary care and screening would require greater investigation to determine the feasibility of such an approach. When the FIT is implemented in the NHS BCSP, SNOMED CT codes and the FIT result could be sent to primary care which could be utilised in research.

The increasing use of EHRs in both clinical practice and research is leading to development of methods to best utilise and extract this data for research. The methods described in **Chapter 6** can be taken forward in other studies to improve quality and validity of data used in this area. Due to the complexity and detail of information available, novel approaches to develop risk prediction models will be required in the future. Machine learning algorithms have shown promise in this research and could be applied for more complex datasets due to their flexibility. Further to this, there are both coded information and free text information stored in EHRs. Methods for free text extraction using text mining techniques warrant investigation in order to utilise this important source of information and can retain anonymity of an individual.

5.0 Conclusions and Recommendations

The research in this thesis has identified the potential of risk-adjusted screening using the FIT for making screening referral decisions. Several growing trends have been identified related to this research area including the increased number of studies in prediction model research, and the growing use of electronic health records as well as the reproducibility and transparency of methods in this area. Machine learning approaches are abundant in other fields of research but tend to be replaced by more conventional methods in health research due to their supposed 'black box' nature. The reporting and quality of risk prediction model studies is often cited as low quality; this was evident in the systematic review in **Chapter 2** where most studies had high risks of bias, particularly in relation to statistical analysis. This is set to improve with the publication of guidelines for developing and reporting models. This thesis endeavored to report methods and model development fully and to provide risk equations by following the TRIPOD guidelines and PROGRESS principles.[10 35]

A relatively simple model which combines routinely available predictors available from the BCSS has shown a significant improvement above just applying the FIT on its own. Before such an approach can be implemented in practice, the model will need to consider further predictors available from the BCSS, externally validate the model in another dataset and

carry out an impact study to assess the effects on patient outcomes. A clear gap in prediction model research identified from the systematic review in **Chapter 2** and from the literature is the lack of external validation studies and, even rarer, impact studies for risk prediction models. Applying a prediction model is essentially applying a new technology and as such the spectrum and characteristics of the diagnosed disease may differ. This needs to be investigated before applying a risk prediction model in practice, despite many models being implemented with a lack of evidence behind their use in guidelines.[31] Future risk prediction model research projects should therefore consider the whole pathway from development to impact, ensuring a separate dataset for external validation and ideally contacting/collaborating with external researchers to carry this out (reduction in bias) as well as planning the assessment of model impact.

Machine learning approaches have shown improvement in performance above and beyond the usual standard statistical methodology. Provided model development is clearly reported and transparent to other researchers to validate results, these methods warrant further investigation. The growing use of EHRs and their more detailed and complex nature may mean machine learning methods are better suited to modelling this data. Future prediction model guidelines should consider reporting of machine learning models and their development. Researchers should ensure methodological transparency of the models and investigate methods to make these approaches more accessible to users (e.g. visualization tools) to improve the adoption of machine learning approaches in practice and to allow external validation by other research groups.[18 47]

Lifestyle, lab parameters, symptoms and other conditions have been identified as predictors, which help to explain an individual's level of risk for colorectal cancer in **Chapter 5** and could be considered for inclusion in a future risk adjusted screening model. The importance of such factors could be considered by offering routine blood tests perhaps as part of health checks or drawing further information from GP records onto the screening database for use in a pre-selection algorithm. It has been suggested that predicting future risk is more valuable than assessing current risk of disease. This can be achieved in a screening programme by assessing an individual's risk at a baseline point and tailoring aspects of screening based on this result. For instance, risk based screening intervals have been discussed in FIT research groups and involve lengthening or shortening an individual's screening interval based on their risk. The 'intended moment' of using a risk prediction

model can be investigated at different points along the screening pathway as an approach to personalize screening.

With the implementation of the FIT by the end of 2018 in the BCSP, there are many further opportunities for research and investigating a risk adjusted approach to screening with the end goal of improving early detection of colorectal cancer and ultimately patient outcomes.

6.0 References

1. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. The British journal of general practice : the journal of the Royal College of General Practitioners. 2012;62(594):e29-37.
2. Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (Colorectal). Br J Cancer. 2012;107(2):260-5.
3. Cleves M, Gould W, Marchenko YV. An Introduction to Survival Analysis Using Stata. Revised third edition ed. Texas, USA: Stata Press; 2016.
4. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis - an introduction to concepts and methods. Br J Cancer. 2003;89(3):431-6.
5. Young GP, Symonds EL, Allison JE, Cole SR, Fraser CG, Halloran SP, et al. Advances in Fecal Occult Blood Tests: The FIT Revolution. Digestive Diseases and Sciences. 2015;60(3):609-22.
6. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. Cancer. 2001;91(8 Suppl):1636-42.
7. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002;35(5-6):352-9.
8. Ahmed FE. Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular Cancer. 2005;4:29-.
9. Biglarian A, Bakhshi E, Gohari MR, Khodabakhshi R. Artificial neural network for prediction of distant metastasis in colorectal cancer. Asian Pacific journal of cancer prevention : APJCP. 2012;13(3):927-30.
10. Moons KM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. Annals of Internal Medicine. 2015;162(1):W1-W73.
11. Hamilton W, Lancashire R, Sharp D, Peters TJ, Cheng K, Marshall T. The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. BMC medicine. 2009;7:17.
12. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS medicine. 2015;12(10):e1001885.
13. van Vugt HA, Roobol MJ, Venderbos LD, Joosten-van Zwanenburg E, Essink-Bot ML, Steyerberg EW, et al. Informed decision making on PSA testing for the detection of prostate cancer: an evaluation of a leaflet with risk indicator. European journal of cancer (Oxford, England : 1990). 2010;46(3):669-77.
14. Edwards AG, Naik G, Ahmed H, Elwyn GJ, Pickles T, Hood K, et al. Personalised risk communication for informed decision making about taking screening tests. The Cochrane database of systematic reviews. 2013(2):Cd001865.
15. Berner ES, Ozaydin B. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2353-4.
16. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. Biometrical journal Biometrische Zeitschrift. 2014;56(4):588-93.
17. Cabitza F, Rasoini R, Gensini G. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8.
18. Huesch MD. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2355-6.
19. Kruppa J, Liu Y, Biau G, Kohler M, Konig IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. Biometrical journal Biometrische Zeitschrift. 2014;56(4):534-63.

20. Kruppa J, Liu Y, Diener HC, Holste T, Weimar C, Konig IR, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: applications. Biometrical journal Biometrische Zeitschrift. 2014;56(4):564-83.
21. McDonald L, Ramagopalan SV, Cox AP, Oguz M. Unintended consequences of machine learning in medicine? F1000Research. 2017;6:1707-.
22. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. Biometrical journal Biometrische Zeitschrift. 2014;56(4):601-6.
23. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. Cancer. 2001;91(8 Suppl):1615-35.
24. Boursi B, Mamtani R, Hwang WT, Haynes K, Yang YX. A Risk Prediction Model for Sporadic CRC Based on Routine Lab Results. Digestive diseases and sciences. 2016;61(7):2076-86.
25. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, et al. Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. Journal of the American Medical Informatics Association : JAMIA. 2016;23(5):879-90.
26. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis - choosing a model and assessing its adequacy and fit. Br J Cancer. 2003;89(4):605-11.
27. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14:40-.
28. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology. 2013;13(1):33.
29. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. First Edition ed. Texas, USA: Stata Press; 2011.
30. Wyatt JC, Altman DG. Commentary: Prognostic models: clinically useful or quickly forgotten? BMJ (Clinical research ed). 1995;311(7019):1539-41.
31. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2):e1001381.
32. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and Methods in Clinical Prediction Research: A Systematic Review. PLoS medicine. 2012;9(5):e1001221.
33. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ (Clinical research ed). 2009;338.
34. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS medicine. 2014;11(10):e1001744.
35. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ (Clinical research ed). 2013;346:e5595.
36. Vickers AJ, Basch E, Kattan MW. Against diagnosis. Ann Intern Med. 2008;149(3):200-3.
37. Chen LS, Yen AM, Chiu SY, Liao CS, Chen HH. Baseline faecal occult blood concentration as a predictor of incident colorectal neoplasia: longitudinal follow-up of a Taiwanese population-based colorectal cancer screening cohort. The Lancet Oncology. 2011;12(6):551-8.

38. Digby J, Fraser CG, Carey FA, Diament RH, Balsitis M, Steele RJ. Faecal haemoglobin concentration is related to detection of advanced colorectal neoplasia in the next screening round. Journal of medical screening. 2017;24(2):62-8.
39. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ (Clinical research ed). 2006;332(7549):1089.
40. Aniwan S, Rerknimitr R, Kongkam P, Wisedopas N, Ponuthai Y, Chaithongrat S, et al. A combination of clinical risk stratification and fecal immunochemical test results to prioritize colonoscopy screening in asymptomatic participants. Gastrointestinal Endoscopy. 2015;81(3):719-27.
41. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175-97.
42. Ng R, Kornas K, Sutradhar R, Wodchis WP, Rosella LC. The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review. Diagnostic and Prognostic Research. 2018;2(1):4.
43. Welten M, de Kroon MLA, Renders CM, Steyerberg EW, Raat H, Twisk JWR, et al. Repeatedly measured predictors: a comparison of methods for prediction modeling. Diagnostic and Prognostic Research. 2018;2(1):5.
44. Good NM, Suresh K, Young GP, Lockett TJ, Macrae FA, Taylor JMG. A prediction model for colon cancer surveillance data. Statistics in medicine. 2015;34(18):2662-75.
45. Carpenter GA, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing. 1987;37(1):54-115.
46. Carpenter GA, Grossberg S, Reynolds JH. ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks. 1991;4(5):565-88.
47. Licitra L, Trama A, Hosni H. Benefits and risks of machine learning decision support systems. JAMA. 2017;318(23):2354-.